VISIT…

LANZAROTE
Caliente.com

ENCYCLOPEDIA OF
MOBILE COMPUTING AND COMMERCE
VOLUME I
David Taniar
Taniar
Encyclopedia of Mobile Computing and Commerce
VOLUME I
VOLUME II
Idea Group Reference

Encyclopedia of Mobile Computing and Commerce

David Taniar
Monash University, Australia

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · London · Melbourne · Singapore

Acquisitions Editor: Kristin Klinger
Development Editor: Kristin Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Sara Reed
Assistant Managing Editor: Diane Huskinson
Copy Editor: Maria Boyer and Alana Bubnis
Typesetter: Diane Huskinson
Support Staff: Sharon Berger, Mike Brehm, Elizabeth Duke, and Jamie Snavely
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of Idea Group Inc.)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.idea-group-ref.com

and in the United Kingdom by
Information Science Reference (an imprint of Idea Group Inc.)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanonline.com

Copyright © 2007 by Idea Group Inc. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of mobile computing and commerce / David Taniar, editor.
p. cm.
Summary: “Nowadays, mobile communication, mobile devices, and mobile computing are widely available. The availability of mobile communication networks has made a huge impact to various applications, including commerce. Consequently, there is a strong relationship between mobile computing and commerce. This book brings to readers articles covering a wide range of mobile technologies and their applications”--Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-59904-002-8 (hardcover) -- ISBN 978-1-59904-003-5 (ebook)
1. Mobile computing--Encyclopedias. 2. Mobile communication systems--Encyclopedias. 3. Mobile commerce--Encyclopedias. I. Taniar, David.
QA76.59.E47 2007
004.16503--dc22
2006039745

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

Editorial Advisory Board

Bernady O. Apduhan
Kyushu Sangyo University, Japan

Irfan Awan
University of Bradford, UK

Leonard Barolli
Fukuoka Institute of Technology, Japan

Stephane Bressan
National University of Singapore, Singapore

Kuo-Ming Chao
Coventry University, UK

Mieso Kabeto Denko
University of Guelph, Canada

Mustafa M. Deris
Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

Arjan Durresi
Louisiana State University, USA

John Goh (Assistant Editor-in-Chief)
Monash University, Australia

D. Frank Hsu
Fordham University, USA

Hai Jin
Huazhong University of Science and Technology, China

Kevin H. Liu
EMC Corporation, USA

Ismail Khalil Ibrahim
Johannes Kepler University Linz, Austria

Jianhua Ma
Hosei University, Japan

Zakaria Maamar
Zayed University, UAE

Joseph K. Ng
Hong Kong Baptist University, Hong Kong

Wenny Rahayu
La Trobe University, Australia

Elhadi Shakshuki
Acadia University, Canada

Timothy K. Shih
Tamkang University, Taiwan

Nguyen Manh Tho
Vienna University of Technology, Austria

Laurence T. Yang
St. Francis Xavier University, Canada

Muhammad Younas
Oxford Brookes University, UK

List of Contributors

Abdul-Mehdi, Ziyad Tariq / *Multimedia University, Malaysia* 233, 369, 693, 947
Ahmad, Ashraf M. A. / *National Chiao Tung University, Taiwan* 627
Al Kattan, Ibrahim / *American University of Sharjah, UAE* 682
Alam, Muhammad Tanvir / *Bond University, Australia* 724, 778
Aleksy, Markus / *University of Mannheim, Germany* 160, 744
Alexiou, Antonios / *Patras University, Greece* 20
Al-Khalifa, Hend S. / *Southampton University, UK* 569
Almeida, Hyggo / *Federal University of Campina Grande, Brazil* 71, 249, 260, 341, 621, 877, 974, 978, 1011
Al-Salman, AbdulMalik S. / *King Saud University, Saudi Arabia* 569
Amara-Hachmi, Nejla / *University of Paris 13, France* 717
Anagnostopoulos, Christos / *University of Athens, Greece* 856
Antonellis, Dimitrios / *Research Academic Computer Technology Institute, Greece & University of Patras, Greece* 20
Antonellis, Ioannis / *University of Patras, Greece* 119
Arbaiy, Nureize / *Kolej Universiti Teknologi Tun Hussein Onn, Malaysia* 763
Avola, Danilo / *Instituto di Ricerche Sulla Popolazione e le Politiche Sociali, Italy* 1050
Aziz Basi, Hussein M. / *Multimedia University, Malaysia* 369, 446
Azzuhri, Saaidal Razalli Bin / *Malaysia University of Science and Technology, Malaysia* 253
Baba, Takaaki / *Waseda University, Japan* 804, 820
Bakhouya, M. / *The George Washington University, Washington DC, USA* 954
Bandyopadhyay, Subir K. / *Indiana University Northwest, USA* 32
Barbosa, Nadia / *Federal University of Campina Grande, Brazil* 877
Basole, Rahul C. / *Georgia Institute of Technology, USA* 481
Beer, Martin / *Sheffield Hallam University, UK* 528
Belsis, Meletis / *Telecron, Greece* 1028
Bin Mamat, Ali / *Universiti Putra Malaysia, Malaysia* 233, 693
Bodomo, Adams / *University of Hong Kong, Hong Kong* 562
Bokor, László / *Budapest University of Technology and Economics, Hungary* 51
Bose, Indranil / *University of Hong Kong, Hong Kong* 96
Bouras, Christos / *Research Academic Computer Technology Institute, Greece & University of Patras, Greece* 20, 119
Bradley, John F. / *University College Dublin, Ireland* 243
Bryant, Barrett R. / *University of Alabama at Birmingham, USA* 436
Bublitz, Frederico / *Federal University of Campina Grande, Brazil* 877
Byrne, Caroline / *Institute of Technology Carlow, Ireland* 310
Carroll, John M. / *The Pennsylvania State University, USA* 291
Caschera, Maria Chiara / *Consiglio Nazionale delle Ricerche, Italy* 675, 1050
Chalmers, Kevin / *Napier University, Scotland* 576
Chan, Alvin T. S. / *The Hong Kong Polytechnic University, Hong Kong* 749
Chand, Narottam / *Indian Institute of Technology Roorkee, India* 172
Chang, Jun-Yang / *National Kaohsiung University of Applied Sciences, Taiwan* 352, 616
Chang, Elizabeth / *Curtin University of Technology, Australia* 108

Chen, Jengchung V. / *National Cheng Kung University, Taiwan* 894
Chen, Shuping / *Beijing University of Posts & Telecommunications, China* 940
Chetan, Kumar S / *NetDevices India Pvt Ltd., India* 195
Chin, Choong Ming / *British Telecommunications (Asian Research Center), Malaysia* 424, 700, 729
Ching-Bin Tse, Alan / *The Chinese University of Hong Kong, Hong Kong* 283
Cho, Vincent / *Hong Kong Polytechnic University, Hong Kong* 38
Chochliouros, Stergios P. / *Hellenic Telecommunications Organization S.A., Greece* 581
Chochliouros, Ioannis P. / *Hellenic Telecommunications Organization S.A., Greece* 581
Chow, Chi-Yin / *University of Minnesota – Twin Cities, USA* 749
Chuang, Li-Yeh / *I-Shou University, Taiwan*....352, 616
Correia, Eduardo / *Christchurch Polytechnic Institute of Technology, New Zealand* 996
Crowther, Paul / *Sheffield Hallam University, UK* 528
Curran, Kevin / *University of Ulster, Northern Ireland* 265, 1022
Cycon, Hans L. / *FHTW Berlin, Germany* 589
da Cunha Borelli, Walter / *State University of Campinas, Brazil* 272
Dagiuklas, Tasos / *Technical Institute of Messolongh, Greece* 357, 796
Dananjayan, P. / *Pondicherry Engineering College, India* 149, 810
da Silva Oliveira, Elthon Allex / *Federal University of Alagoas – Campus Arapiraca, Brazil* 987
de Araújo Lima, Emerson Ferreira / *Federal University of Campina Grande, Brazil* 987
de Carvalho Gomes, Yuri / *Federal University of Campina Grande, Brazil* 1011
de Figueiredo, Jorge César Abrantes / *Federal University of Campina Grande, Brazil* 987
de Leoni, Massimiliano / *University of Rome "La Sapienza", Italy* 1043
de Oliveira, Juliano Rodrigues Fernandes / *Federal University of Campina Grande, Brazil* 1011
De Rosa, Fabio / *University of Rome "La Sapienza", Italy* 1043
Decker, Michael / *University of Karlsruhe, Germany* 398, 711
Denko, Mieso Kabeto / *University of Guelph, Canada* 328
Deris, Mustafa M. / *Kolej Universiti Teknologi Tun Hussein Onn, Malaysia* 763
Di Noia, Tommaso / *Politecnico di Bari, Italy* 43
Di Sciascio, Eugenio / *Politecnico di Bari, Italy* 43
Diekmann, Thomas / *University of Goettingen, Germany* 124
Dillon, Tharam S. / *University of Technology, Sydney, Australia* 108
Dirs, Mustafa M. / *College University Technology Tun Hussein Onn, Malaysia* 233, 693, 947
Djordjevic-Kajan, Slobodanka / *University of Nis, Serbia* 129, 660
Damodaran, Dhilak / *Monash University, Australia* 1015
Donini, Francesco Maria / *Università della Tuscia, Italy* 43
Dudás, István / *Budapest University of Technology and Economics, Hungary* 51
El Fallah-Seghrouchni, Amal / *University of Paris 6, France* 717
El Morr, Christo / *York University, Canada* 632
El-Said, Mostafa / *Grand Valley State University, USA* 63, 688
Ferreira, Glauber / *Federal University of Campina Grande, Brazil* 877
Ferri, Fernando / *Istituto di Richerche Sulla Popolazione e le Politiche Sociali – CNR, Italy* 675, 1050
Fleet, Gregory John / *University of New Brunswick at Saint John, Canada* 78
Flores, Andres / *University of Comahue, Argentina*....59
Freire de Souza Santos, Danilo / *Federal University of Campina Grande, Brazil* 341
Gaber, J. / *Université de Technologie de Belfort-Montbéliard, France* 954
Gandhamaneni, Jayasree / *Indiana University Purdue University Indianapolis, USA* 436
Ganoe, Craig H. / *The Pennsylvania State University, USA*....291
Garcia, Juan / *Illinois State University, USA* 461
García-Macías, J. Antonio / *CICESE Research Center, Mexico* 773
Gardikis, G. / *University of Aegean, Greece* 889
Garret, Bernie / *University of British Columbia, Canada* 754
Goldberg, Steve / *INET International Inc., Canada* 1004
Grahn, Kaj / *Arcada Polytechnic, Finland* 839
Griffiths, Mark / *Nottingham Trent University, UK* 553

Grifoni, Patrizia / *Istituto di Ricerche Sulla Popolazione e le Politiche Sociali – CNR, Italy* 675, 1050
Gritzalis, Stefanos / *University of the Aegean, Greece* 1028
Guan, Jihong / *Tongji University, China* 84, 213, 789
Guan, Sheng-Uei / *Brunel University, UK* 334, 345, 429, 826
Gurău, Călin / *Montpellier Business School, France* 557, 999
Gyasi-Agyei, Amoakoh / *Central Queensland University, Australia* 165
Hadjiefthymiades, Stathes / *University of Athens, Greece* 856, 863
Hagenhoff, Svenja / *Georg-August-University of Goettingen, Germany* 124
Hartung, Frank / *Ericsson GmbH, Germany* 611
Hegedüs, Péter / *Budapest University of Technology and Economics, Hungary* 393
Herbster, Raul Fernandes / *Federal University of Campina Grande, Brazil* 260, 974
Hiew, Pang Leang / *British Telecommunications (Asian Research Center), Malaysia* 487, 906
Hoh, Simon / *British Telecommunications (Asia Research Center), Malaysia* 138
Horn, Uwe / *Ericsson GmbH, Germany* 611
Hosszú, Gábor / *Budapest University of Technology and Economics, Hungary* 393
Hsu, Wen-Jing / *Nanyang Technological University, Singapore* 734
Hu, Wen-Chen / *University of North Dakota, USA* 302
Huang, Bo / *Waseda University, Japan* 804, 820
Huang, Hong / *New Mexico State University, USA* 202
Hung, Humphry / *Hong Kong Polytechnic University, Hong Kong* 38
Hussain, Omar Khadeer / *Curtin University of Technology, Australia* 108
Hussain, Farookh Khadeer / *University of Technology, Australia* 108
Ibrahim, Hamidah / *Universiti Putra Malaysia, Malaysia* 233, 693, 947
Ifinedo, Princely / *University of Jyvaskyla, Finland* 605
Imre, Sándor / *Budapest University of Technology and Economics, Hungary* 51
Iris, Reychav / *Bar-Ilan University, Israel* 413
Jasimuddin, Sajjad M. / *University of Wales – Aberystwyth, UK* 520
Jayaputera, James W. / *Monash University, Australia* 739
Jeong, Eui Jun / *Michigan State University, USA* 185, 928
Jiménez, Leonardo Galicia / *CICESE Research Center, Mexico* 773
Joshi, R. C. / *Indian Institute of Technology Roorkee, India* 172
Ju, Khoo Wei / *Malaysia University of Science and Technology, Malaysia* 912
Kaldanis, Vasileios S. / *NTUA, Greece* 1
Kalliaras, Panagiotis / *National Technical University of Athens, Greece* 381, 387, 960, 981
Kampmann, Markus / *Ericsson GmbH, Germany* 611
Kamthan, Pankaj / *Concordia University, Canada* 9, 25, 277, 375,
Kao, I-Lung / *IBM, USA* 302
Karlsson, Jonny / *Arcada Polytechnic, Finland* 839
Karnouskos, Stamatis / *SAP AG, Germany* 706
Kartham, Pankaj / *Concordia University, Canada* 9, 25, 277
Kaspar, Christian / *University of Goettingen, Germany* 124
Katsukura, Akihisa / *Dentsu Inc., Japan* 639
Keegan, Stephen / *University College Dublin, Ireland* 310
Kennedy, David M. / *Hong Kong University, Hong Kong* 317
Kerridge, Jon / *Napier University, Scotland* 576
Khashchanskiy, Victor / *First Hop Ltd., Finland* 15, 785
Kim, Dan J. / *University of Houston Clear Lake, USA* 185, 928
Kini, Ranjan B. / *Indiana University Northwest, USA* 32
Kitisin, Sukumal / *Kasetsart University, Thailand* 220
Kleinschmidt, João Henrique / *State University of Campinas, Brazil* 272
Korhonen, Jouni / *TeliaSonera Corporation, Finland* 966
Korthaus, Axel / *University of Mannheim, Germany* 160
Kotsopoulos, Stavros / *University of Patras, Greece* 357, 796
Koukia, Spiridoula / *Universtiy of Greece, Greece* 116

Koumaras, H. / *N.C.S.R., Demokritos, Greece* 758, 889
Kourtis, A. / *N.C.S.R., Demokritos, Greece* 758, 889
Kovács, Ferenc / *Budapest University of Technology and Economics, Hungary* 393
Kritzner, Jan / *Aachen University, Germany* 611
Kustov, Andrei L. / *First Hop Ltd., Finland* 15, 785
Kvasnica, Milan / *Tomas Bata University, Zlin, Czech Republic* 403, 651
Lalopoulos, George K. / *Hellenic Telecommunications Organization S.A., Greece* 581
Lang, Jia / *Nice Business Solutions Finland, Finland* 785
Lau, Chiew-Tong / *Nanyang Technological University, Singapore* 734
Le, Phu Dung / *Monash University, Australia* 227, 832, 1015
Lee, Cheon-Pyo / *Carson-Newman College, USA* 442
Lee, Dennis / *The University of Queensland, Australia & The Australian CRC for Interactive Design, Australia* 933
Lei, Pouwan / *University of Bradford, UK* 455
Leong, Hong Va / *The Hong Kong Polytechnic University, Hong Kong* 749
Leu, Huei / *Industrial Technology Research Institute, Taiwan* 178
Liberati, Diego / *Italian National Research Council, Italy* 68
Lim, Say Ying / *Monash University, Australia* 102, 154, 849
Lin, Chad / *Edith Cowan University, Australia* 178
Lin, Koong / *Taiwan National University of the Arts, Taiwan* 178
Liu, Chao / *Waseda University, Japan* 804, 820
Lívio Vasconcelos Guedes, Ádrian / *Federal University of Campina Grande, Brazil* 249
Lonthoff, Jörg / *Technische Universität Darmstadt, Germany* 510
Loureiro, Emerson / *Federal University of Campina Grande, Brazil* 71, 877
Luís do Nascimento, José / *Federal University of Campina Grande, Brazil* 341
Maamar, Zakaria / *Zayed University, UAE* 190
Mahatanankoon, Pruthikrai / *Illinois State University, USA* 461
Mahmoud, Qusay H. / *University of Guelph, Canada* 190
Malik, Haroon / *Acadia University, Canada* 328
Mamat, Ali Bin / *FSKTM – UPM, Malaysia* 693, 947
Maricar, Habeebur Rahman / *American University of Sharjah, UAE* 682
Marques, Stefânia / *Federal University of Campina Grande, Brazil* 978
Martakos, D. / *University of Athens, Greece* 758
Massimi, Michael / *University of Toronto, Canada* 291
McMeel, Dermott / *University of Edinburgh, Scotland* 516
Mecella, Massimo / *University of Rome, Italy* 1043
Menipaz, Ehud / *Ben-Gurion University, Israel* 413
Merten, Patrick S. / *University of Fribourg, Switzerland* 466
Misra, Manoj / *Indian Institute of Technology Roorkee, India* 172
Mitchell, Stella / *IBM T. J. Watson Research, USA* 644
Morais, Marcos / *Federal University of Campina Grande, Brazil* 260
Muhlberger, Ralf / *The University of Queensland, Australia & The Australian CRC for Interactive Design, Australia* 933
Muldoon, Conor / *University College Dublin, Ireland* 243
Nanopoulos, Alexandros / *Aristotle University, Greece* 660
Nishiyama, Mamoru / *Dentsu Communication Institute Inc., Japan* 639
O'Grady, Michael J. / *University College Dublin, Ireland* 243, 769, 1034
O'Hare, Gregory M. P. / *University College Dublin, Ireland* 243, 310, 769, 1034
O'Hare, Peter / *University College Dublin, Ireland* 310
Okazaki, Shintaro / *Autonomous University of Madrid, Spain* 296, 635, 639, 885
Oliveira, Loreno / *Federal University of Campina Grande, Brazil* 71, 621, 877
Olla, Phillip / *Madonna University, USA* 504
Olson, Andrew M. / *Indiana University Purdue University Indianapolis, USA* 436
Orosz, Mihály / *Budapest University of Technology and Economics, Hungary* 393

Pallis, E. / *Technological Educational Institute of Crete, Greece* 758, 889
Papadopoulos, Apostolos N. / *Aristotle University, Greece* 660
Papageorgiou, P. / *National Technical University of Athens, Greece* 387
Parsons, David / *Massey University, New Zealand* 525
Patel, Keyurkumar J. / *Box Hill Institute, Australia* 365
Patel, Umesh / *Box Hill Institute, Australia* 365
Patrikakis, Charalampos Z. / *NTUA, Greece* 1
Paulo de Assis Barbosa, Luiz / *Federal University of Campina Grande, Brazil* 1011
Pavlovski, Christopher J. / *IBM Corporation, Australia* 644, 870
Peng, Mugen / *Beijing University of Posts & Telecommunications, China* 921, 940
Perkusich, Angelo / *Federal University of Campina Grande, Brazil* 71, 249, 260, 341, 621, 877, 974, 978, 1011
Petrova, Krassie / *Auckland University of Technology, New Zealand* 497, 899
Piscitelli, Giacomo / *Politecnico di Bari, Italy* 43
Plant, Laurence / *IBM Corporation, Australia* 870
Politis, Ilias / *University of Patras, Greece* 357, 796
Poulopoulos, Vassilis / *Research Academic Computer Technology Institute, Greece & University of Patras, Greece* 119
Pousttchi, Key / *University of Augsburg, Germany* 547
Predić, Bratislav / *University of Nis, Serbia* 129
Protonotarios, Vasileios E. / *NTUA, Greece* 1
Pulkkis, Göran / *Arcada Polytechnic, Finland* 839
Queiroga, Miguel / *Federal University of Campina Grande, Brazil* 978
Raisinghani, Mahesh S. / *TWU School of Management, USA* 472
Raje, Rajeev R. / *Indiana University Purdue University Indianapolis, USA* 207, 436
Ramamurthy, M. B. / *Multimedia University, Malaysia* 446
Ramli, Azizul Azhar / *Kolej Universiti Teknologi Tun Hussein Onn, Malaysia* 763
Reid, Jeffery G. / *xwave Saint John, Canada* 78
Rigou, Maria / *University of Patras, Greece & Research Academic Computer Technology Institute, Greece* 116
Romdhami, Imed / *Napier University, Scotland* 576
Rouse, William B / *Georgia Institute of Technology, USA* 481
Ruta, Michele / *Politecnico di Bari, Italy* 43
Ruzzelli, Antonio G. / *University College Dublin, Ireland* 1034
Saravanan, I. / *Pondicherry Engineering College, India* 149, 810
Schader, Martin / *University of Mannheim, Germany* 160, 744
Schmidt, Thomas C. / *HAW Hamburg, Germany* 541, 589
Seet, Boon-Chong / *Nanyang Technological University, Singapore* 734
Sekkas, Odysseas / *University of Athens, Greece* 863
Serenko, Alexander / *Lakehead University, Canada* 143
Shakshuki, Elhadi / *Acadia University, Canada* 328
Shirali-Shahreza, Mohammad / *Sharif University of Technology, Iran* 666
Silva Rocha, Jerônimo / *Federal University of Campina Grande, Brazil* 249
Sim, Moh Lim / *Multimedia University, Malaysia* 424, 700, 729
Simitsis, Alkis / *National Technical University of Athens, Greece* 1028
Singh, Rohit / *Monash University, Australia* 1015
Sircar, Ranapratap / *Wipro Technologies, India* 195
Sirmakessis, Spiros / *Technological Institution of Messolongi & Research Academic Computer Technology Institute, Greece* 116
Sivaradje, G. / *Pondicherry Engineering College, India* 149, 810
Smyth, Elaine / *University of Ulster, Northern Ireland* 1022
So, Simon / *Hong Kong Institute of Education, Hong Kong* 419
Sotiriou, Athanasios-Dimitrios / *National Technical University of Athens, Greece* 381, 387, 960, 981
Souto, Sabrina / *Federal University of Campina Grande, Brazil* 978
Spiliopoulou, Anastasia S. / *Hellenic Telecommunications Organization S.A., Greece* 581
Srikhutkhao, Nopparat / *Kasetsart University, Thailand* 220

Steinert, Martin / *University of Fribourg, Switzerland* ... 466

Stojanović, Dragan / *University of Nis, Serbia* ... 129, 660

Suradi, Zurinah / *Kolej Universiti Teknologi Tun Hussein Onn, Malaysia* ... 763

Tan, Chor Min / *British Telecommunications (Asian Research Center), Malaysia* ... 424, 700, 729, 906

Tarkoma, Sasu / *Helsinki Institute for Information Technology, Finland* ... 966

Tay, Yuan Sherng / *National University of Singapore, Singapore* ... 345

Teufel, Stephanie / *University of Fribourg, Switzerland* ... 466

Tjondronegoro, Dian / *Queensland University of Technology, Australia* ... 596

Tong, Carrison K. S. / *Pamela Youde Nethersole Eastern Hospital, Hong Kong* ... 533

Tran, Dai / *Arcada Polytechnic, Finland* ... 839

Tsagkaropoulos, Michail / *University of Patras, Greece* ... 357, 796

Tsetsos, Vassileios / *University of Athens, Greece* ... 856

Tuceryan, Mihran / *Indiana University Purdue University Indianapolis, USA* ... 207

Turel, Ofir / *McMaster University, Canada* ... 143

Tynan, Richard / *University College Dublin, Ireland* ... 1034

Usaola, Macario Polo / *Universidad de Castilla-La Mancha, Spain* ... 57

Venkataram, P. / *Indian Institute of Science, India* ... 195

Vogel, Doug / *City University of Hong Kong, Hong Kong* ... 317

Wählisch, Matthias / *FHTW Berlin, Germany* ... 541, 589

Wang, Yiling / *Monash University, Australia* ... 227, 832

Wang, JiaJia / *University of Bradford, UK* ... 455

Wang, Laura / *Tongji University, China* ... 669

Wang, Yiling / *Monash University, Australia* ... 227, 832

Wang, Wenbo / *Beijing University of Posts & Telecommunications, China* ... 921, 940

Wang, Yingjie / *Beijing University of Posts & Telecommunications, China* ... 921

Wickramasinghe, Nilmini / *IIllinois Institute of Technology, USA* ... 1004

Wiedemann, Dietmar Georg / *University of Augsburg, Germany* ... 547

Willis, Robert / *Lakehead University, Canada* ... 143

Wong, Chin Chin / *British Telecommunications (Asian Research Center), Malaysia* ... 138, 487, 906

Wong, K. Daniel / *Malaysia University of Science and Technology, Malaysia* ... 253, 912

Wong, King Yin / *The Chinese University of Hong Kong, Hong Kong* ... 283

Wong, Eric T. T. / *Hong Kong Polytechnic University, Hong Kong* ... 533

Wright, David / *University of Ottawa, Canada* ... 90, 816, 1038

Xavier, Rodrigo Nóbrega Rocha / *Federal University of Campina Grande, Brazil* ... 1011

Xi, Chen / *University of Hong Kong, Hong Kong* ... 96

Xilouris, G. / *N.C.S.R., Demokritos, Greece* ... 758, 889

Yan, Lu / *Åbo Akademi, Finland* ... 492

Yan, Hong / *City University of Hong Kong, Hong Kong & University of Sydney, Australia* ... 669

Yang, Cheng-Hong / *National Kaohsiung University of Applied Sciences, Taiwan* ... 352, 616

Yang, Cheng Huei / *National Kaohsiung Marine University, Taiwan* ... 352, 616

Ye, Yang / *Tongji University, China* ... 669

Yeh, Jyh-haw / *Boise State University, USA* ... 302

Yim, Frederick Hong Kit / *Drexel University, USA* ... 283

Zervas, Evangelos / *Tei-Athens, Greece* ... 863

Zhang, Zuopeng (Justin) / *Eastern New Mexico University, USA* ... 520

Zhong, Yapin / *Shandog Institute of Physical Education and Sport, China* ... 302

Zhou, Jiaogen / *Wuhan University, China* ... 84, 213, 789

Zhou, Shuigeng / *Fudan University, China* ... 84, 213, 789

Zhu, Fubao / *Wuhan University, China* ... 213, 789

Zoi, S. / *National Technical University of Athens, Greece* ... 387

Contents
by Volume

VOLUME I

Academic Activities Based on Personal Networks Deployment / *Vasileios S. Kaldanis, Charalampos Z. Patrikakis, and Vasileios E. Protonotarios* 1

Accessibility of Mobile Applications / *Pankaj Kamthan* 9

Acoustic Data Communication with Mobile Devices / *Victor I. Khashchanskiy and Andrei L. Kustov* 15

Adaptive Transmission of Multimedia Data over UMTS / *Antonios Alexiou, Dimitrios Antonellis, and Christos Bouras* 20

Addressing the Credibility of Mobile Applications / *Pankaj Kamthan* 25

Adoption and Diffusion of M-Commerce / *Ranjan B. Kini and Subir K. Bandyopadhyay* 32

Adoption of M-Commerce Devices by Consumers / *Humphry Hung and Vincent Cho* 38

Advanced Resource Discovery Protocol for Semantic-Enabled M-Commerce / *Michele Ruta, Tommaso Di Noia, Eugenio Di Sciascio, Francesco Maria Donini, and Giacomo Piscitelli* 43

Anycast-Based Mobility / *István Dudás, László Bokor, and Sándor Imre* 51

Applications Suitability on PvC Environments / *Andres Flores and Macario Polo Usaola* 57

Bio-Inspired Approach for the Next Generation of Cellular Systems, A / *Mostafa El-Said* 63

Brain Computer Interfacing / *Diego Liberati* 68

Bridging Together Mobile and Service Oriented Computing / *Loreno Oliveira, Emerson Loureiro, Hyggo Almeida, and Angelo Perkusich* 71

Browser-Less Surfing and Mobile Internet Access / *Gregory John Fleet and Jeffery G. Reid* 78

Building Web Services in P2P Networks / *Jihong Guan, Shuigeng Zhou, and Jiaogen Zhou* 84

Business and Technology Issues in Wireless Networking / *David Wright* 90

Business Strategies for Mobile Marketing / *Indranil Bose and Chen Xi* 96

Cache Invalidation in a Mobile Environment / *Say Ying Lim* 102

Communicating Recommendations in a Service-Oriented Environment / *Omar Khadeer Hussain, Elizabeth Chang, Farookh Khadeer Hussain, and Tharam S. Dillon* 108

Content Personalization for Mobile Interfaces / *Spiridoula Koukia, Maria Rigou, and Spiros Sirmakessis* 116

Content Transformation Techniques / *Ioannis Antonellis, Christos Bouras, and Vassilis Poulopoulos* 119

Context-Adaptive Mobile Systems / *Christian Kaspar, Thomas Diekmann, and Svenja Hagenhoff* 124

Context-Aware Mobile Geographic Information Systems / *Slobodanka Djordjevic-Kajan, Dragan Stojanović, and Bratislav Predić* 129

Context-Aware Systems / *Chin Chin Wong and Simon Hoh* 138

Contractual Obligations between Mobile Service Providers and Users / *Robert Willis, Alexander Serenko, and Ofir Turel* 143

Convergence Technology for Enabling Technologies / *G. Sivaradje, I. Saravanan, and P. Dananjayan* 149

Cooperative Caching in a Mobile Environment / *Say Ying Lim* 154

CORBA on Mobile Devices / *Markus Aleksy, Axel Korthaus, and Martin Schader* 160

Cross-Layer RRM in Wireless Data Networks / *Amoakoh Gyasi-Agyei* 165

Data Caching in Mobile Ad-Hoc Networks / *Narottam Chand, R. C. Joshi, and Manoj Misra* 172

Decision Analysis for Business to Adopt RFID / *Koong Lin, Chad Lin, and Huei Leu* 178

Definitions, Key Characteristics, and Generations of Mobile Games / *Eui Jun Jeong and Dan J. Kim* 185

Design Methodology for Mobile Information Systems / *Zakaria Maamar and Qusay H. Mahmoud* 190

Distributed Approach for QoS Guarantee to Wireless Multimedia / *Kumar S. Chetan, P. Venkataram, and Ranapratap Sircar* 195

Distributed Computing in Wireless Sensor Networks / *Hong Huang* 202

Distributed Heterogeneous Tracking for Augmented Reality / *Mihran Tuceryan and Rajeev R. Raje* 207

Distributed Web GIS / *Jihong Guan, Shuigeng Zhou, Jiaogen Zhou, and Fubao Zhu* 213

Dynamic Pricing Based on Net Cost for Mobile Content Services / *Nopparat Srikhutkhao, and Sukumal Kitisin* 220

Efficient and Scalable Group Key Management in Wireless Networks / *Yiling Wang and Phu Dung Le* 227

Efficient Replication Management Techniques for Mobile Databases / *Ziyad Tariq Abdul-Mehdi, Ali Bin Mamat, Hamidah Ibrahim, and Mustafa Mat Dirs* 233

Embedded Agents for Mobile Services / *John F. Bradley, Conor Muldoon, Gregory M. P. O'Hare, and Michael J. O'Grady* 243

Enabling Mobile Chat Using Bluetooth / *Ádrian Lívio Vasconcelos Guedes, Jerônimo Silva Rocha, Hyggo Almeida, and Angelo Perkusich* 249

Enabling Mobility in IPv6 Networks / *Saaidal Razalli Bin Azzuhri and K. Daniel Wong* 253

Enabling Multimedia Applications in Memory-Limited Mobile Devices / *Raul Fernandes Herbster, Hyggo Almeida, Angelo Perkusich, and Marcos Morais* 260

Enabling Technologies for Mobile Multimedia / *Kevin Curran* 265

Enabling Technologies for Pervasive Computing / *João Henrique Kleinschmidt and Walter da Cunha Borelli* 272

Extreme Programming for Mobile Applications / *Pankaj Kartham* 277

Factors Affecting Mobile Commerce and Level of Involvement / *Frederick Hong Kit Yim, Alan ching Biu Tse, and King Yin Wong* 283

Game-Based Methodology for Collaborative Mobile Applications, A / *Michael Massimi, Craig H. Ganoe, and John M. Carroll* 291

Gender Difference in the Motivations of Mobile Internet Usage / *Shintaro Okazaki* 296

Handheld Computing and J2ME for Internet-Enabled Mobile Handheld Devices / *Wen-Chen Hu, Jyh-haw Yeh, I-Lung Kao, and Yapin Zhong* 302

Infrastructural Perspective on U-Commerce, An / *Stephen Keegan, Caroline Byrne, Peter O'Hare, and Gregory M. P. O'Hare* 310

Integrating Pedagogy, Infrastructure, and Tools for Mobile Learning / *David M. Kennedy and Doug Vogel* 317

Intelligent Medium Access Control Protocol for WSN / *Haroon Malik, Elhadi Shakshuki, and Mieso Kabeto Denko* 328

Intelligent User Preference Detection for Product Brokering / *Sheng-Uei Guan* 334

Interactive Multimedia File Sharing Using Bluetooth / *Danilo Freire de Souza Santos, José Luís do Nascimento, Hyggo Almeida, and Angelo Perkusich* 341

Interactive Product Catalog for M-Commerce / *Sheng-Uei Guan and Yuan Sherng Tay* 345

Interactive Wireless Morse Code Learning System, An / *Cheng-Huei Yang , Li Yeh Chuang, Cheng-Hong Yang, and Jun-Yang Chang,* 352

Interworking Architectures of 3G and WLAN / *Ilias Politis, Tasos Dagiuklas, Michail Tsagkaropoulos, and Stavros Kotsopoulos* 357

iPod as a Visitor's Personal Guide / *Keyurkumar J. Patel and Umesh Patel* 365

Keyword-Based Language for Mobile Phones Query Services / *Ziyad Tariq Abdul-Mehdi and Hussein M. Aziz Basi* 369

Knowledge Representation in Semantic Mobile Applications / *Pankaj Kamthan* 375

Location-Based Multimedia Content Delivery System for Monitoring Purposes / *Athanasios-Dimitrios Sotiriou and Panagiotis Kalliaras* 381

Location-Based Multimedia Services for Tourists / *P. Kalliaras, Athanasios-Dimitrios Sotiriou, P. Papageorgiou, and S. Zoi* 387

Location-Based Services / *Péter Hegedüs, Mihály Orosz, Gábor Hosszú, and Ferenc Kovács* 393

M-Advertising / *Michael Decker* 398

Man-Machine Interface with Applications in Mobile Robotic Systems / *Milan Kvasnica* 403

M-Commerce Technology Perceptions on Technology Adoptions / *Reychav Iris and Ehud Menipaz* 413

M-Learning with Mobile Phones / *Simon So* 419

Mobile Ad-Hoc Networks / *Moh Lim Sim, Choong Ming Chin, and Chor Min Tan* 424

Mobile Agent Protection for M-Commerce / *Sheng-Uei Guan* 429

Mobile Agent-Based Discovery System / *Rajeev R. Raje, Jayasree Gandhamaneni, Andrew M. Olson, and Barrett R. Bryant* 436

Mobile Business Applications / *Cheon-Pyo Lee* 442

Mobile Cellular Traffic with the Effect of Outage Channels / *Hussein M. Aziz Basi and M. B. Ramamurthy* 446

Mobile Commerce / *JiaJia Wang and Pouwan Lei* 455

Mobile Commerce Adoption Barriers / *Pruthikrai Mahatanankoon and Juan Garcia* 461

Mobile Computing and Commerce Framework, A / *Stephanie Teufel, Patrick S. Merten, and Martin Steinert* 466

Mobile E-Commerce as a Strategic Imperative for the New Economy / *Mahesh S. Raisinghani* 472

Mobile Enterprise Readiness and Transformation / *Rahul C. Basole and William B. Rouse* 481

Mobile Entertainment / *Chin Chin Wong and Pang Leang Hiew* 487

Mobile File-Sharing over P2P Networks / *Lu Yan* 492

Mobile Gaming / *Krassie Petrova* 497

Mobile Healthcare Communication Infrastructure Networks / *Phillip Olla* 504

Mobile Hunters / *Jörg Lonthoff* 510

VOLUME II

Mobile ICT / *Dermott McMeel* 516

Mobile Knowledge Management / *Zuopeng (Justin) Zhang and Sajjad M. Jasimuddin* 520

Mobile Learning / *David Parsons* 525

Mobile Learning Environments / *Paul Crowther and Martin Beer* 528

Mobile Medical Image Viewing Using 3G Wireless Network / *Carrison K. S. Tong and Eric T. T. Wong* 533

Mobile Multicast / *Thomas C. Schmidt and Matthias Wählisch* 541

Mobile Payment and the Charging of Mobile Services / *Key Pousttchi and Dietmar Georg Wiedemann* 547

Mobile Phone Gambling / *Mark Griffiths* 553

Mobile Phone Privacy Issues / *Călin Gurău* 557

Mobile Phone Texting in Hong Kong / *Adams Bodomo* 562

Mobile Phones for People with Disabilities / *Hend S. Al-Khalifa and AbdulMalik S. Al-Salman* 569

Mobile Processes and Mobile Channels / *Kevin Chalmers, Imed Romdhami, and Jon Kerridge* 576

Mobile Public Key Infrastructures / *Ioannis Chochliouros, George K. Lalopoulos, Stergios P. Chochliouros, and Anastasia S. Spiliopoulou* 581

Mobile Serverless Video Communication / *Hans L. Cycon, Thomas C. Schmidt, and Matthias Wählisch* 589

Mobile Sports Video with Total Users Control / *Dian Tjondronegoro* 596

Mobile Telephony in Sub-Saharan Africa / *Princely Ifinedo* 605

Mobile Television / *Frank Hartung, Markus Kampmann, Uwe Horn, and Jan Kritzner* 611

Mobile Text Messaging Interface for Persons with Physical Disabilities / *Cheng-Huei Yang, Li-Yeh Chuang, Cheng-Hong Yang, and Jun-Yang Chang* 616

Mobile Users in Smart Spaces / *Loreno Oliveira, Hyggo Almeida, and Angelo Perkusich* 621

Mobile Video Transcoding Approaches and Challenges / *Ashraf M. A. Ahmad* 627

Mobile Virtual Communities / *Christo El Morr* 632

Mobile-Based Advertising in Japan / *Shintaro Okazaki* 635

Mobile-Based Research Methods / *Shintaro Okazaki, Akihisa Katsukura, and Mamoru Nishiyama* 639

Mobility and Multimodal User Interfaces / *Christopher J. Pavlovski and Stella Mitchell* 644

Modular Sensory System for Robotics and Human-Machine Interaction Based on Optoelectronic Components / *Milan Kvasnica* 651

Monitoring and Tracking Moving Objects in Mobile Environments / *Dragan Stojanovic, Slobodanka Djordjevic-Kajan, Apostolos N. Papadopoulos, and Alexandros Nanopoulos* 660

Multilingual SMS / *Mohammad Shirali-Shahreza* 666

Multimedia Contents for Mobile Entertainment / *Hong Yan, Laura Wang, and Yang Ye* 669

Multimodality in Mobile Applications and Services / *Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni* 675

Multi-User OFDM in Mobile Multimedia Network / *Ibrahim Al Kattan and Habeebur Rahman Maricar* 682

Mutual Biometric Authentication / *Mostafa El-Said* 688

New Transaction Management Model / *Ziyad Tariq Abdul-Mehdi, Ali Bin Mamat, Hamidah Ibrahim, and Mustafa M. Dirs* 693

Next-Generation Mobile Technologies / *Chor Min Tan, Choong Ming Chin, and Moh Lim Sim* 700

NFC-Capable Mobile Devices for Mobile Payment Services / *Stamatis Karnouskos* 706

Notification Services for Mobile Scenarios / *Michael Decker* 711

Ontology-Based Approach for Mobile Agent's Context-Awareness, An / *Nejla Amara-Hachmi and Amal El Fallah-Seghrouchni* 717

Optimal Timer for Push to Talk Controller, An / *Muhammad Tanvir Alam* 724

Optimal Utilisation of Future Wireless Resources / *Choong Ming Chin, Chor Min Tan, and Moh Lim Sim* 729

P2P Models and Complexity in MANETs / *Boon-Chong Seet, Chiew-Tong Lau, and Wen-Jing Hsu* 734

Partial Global Indexing for Location-Dependent Query Processing / *James W. Jayaputera* 739

Patterns for Mobile Applications / *Markus Aleksy and Martin Schader* 744

Peer-to-Peer Cooperative Caching in Mobile Environments / *Chi-Yin Chow, Hong Va Leong, and Alvin T. S. Chan* 749

Pen-Based Mobile Computing / *Bernie Garret* 754

Perceived Quality Evaluation for Multimedia Services / *H. Koumaras, E. Pallis, G. Xilouris, A. Kourtis, and D. Martakos* 758

Pest Activity Prognosis in the Rice Field / *Nureize Arbaiy, Azizul Azhar Ramli, Zurinah Suradi, and Mustafa Mat Deris* 763

Positioning Technologies for Mobile Computing / *Michael J. O'Grady and Gregory O'Hare* 769

Privacy Concerns for Indoor Location-Based Services / *Leonardo Galicia Jiménez, and J. Antonio García-Macías* 773

Protocol Analysis for the 3G IP Multimedia Subsystem / *Muhammad Tanvir Alam* 778

Protocol Replacement Proxy for 2.5 and 3G Mobile Internet / *Victor Khashchanskiy, Andrei Kustov, and Jia Lang* 785

Providing Location-Based Services under Web Services Framework / *Jihong Guan, Shuigeng Zhou, Jiaogen Zhou, and Fubao Zhu* 789

Provisioning of Multimedia Applications across Heterogeneous All-IP Networks / *Michail Tsagkaropoulos, Ilias Politis, Tasos Dagiuklas, and Stavros Kotsopoulos* 796

QoS Routing Framework on Bluetooth Networking, A / *Chao Liu, Bo Huang, and Takaaki Baba* 804

Radio Resource Management in Convergence Technologies / *G. Sivaradje, I. Saravanan, and P. Dananjayan* 810

RFID and Wireless Personal Area Networks for Supply Chain Management / *David Wright* 816

Scatternet Structure for Improving Routing and Communication Performance / *Bo Huang, Chao Liu, and Takaaki Baba* 820

Secure Agent Data Protection for E-Commerce Applications / *Sheng-Uei Guan* 826

Secure Group Communications in Wireless Networks / *Yiling Wang and Phu Dung Le* 832

Security Architectures of Mobile Computing / *Kaj Grahn, Göran Pulkkis, Jonny Karlsson, and Dai Tran* 839

Semantic Caching in a Mobile Environment / *Say Ying Lim* 849

Semantic Enrichment of Location-Based Services / *Vassileios Tsetsos, Christos Anagnostopoulos, and Stathes Hadjiefthymiades* 856

Sensor Data Fusion for Location Awareness / *Odysseas Sekkas, Stathes Hadjiefthymiades, and Evangelos Zervas* 863

Service Delivery Platforms in Mobile Convergence / *Christopher J. Pavlovski and Laurence Plant* 870

Service Provision for Pervasive Computing Environments / *Emerson Loureiro, Frederico Bublitz, Loreno Oliveira, Nadia Barbosa, Angelo Perkusich, Hyggo Almeida, and Glauber Ferreira* 877

Short Message Service (SMS) as an Advertising Medium / *Shintaro Okazaki* 885

Shot Boundary Detection Techniques for Video Sequences / *H. Koumaras, G. Xilouris, E. Pallis, G. Gardikis, and A. Kourtis* 889

Smartphone Acceptance among Sales Drivers / *Jengchung V. Chen* 894

SMS-Based Mobile Learning / *Krassie Petrova* 899

Snapshot Assessment of Asia Pacific BWA Business Scenario / *Chin Chin Wong, Chor Min Tan, and Pang Leang Hiew* 906

Software Platforms for Mobile Programming / *Khoo Wei Ju and K. Daniel Wong* 912

Standard-Based Wireless Mesh Networks / *Mugen Peng, Yingjie Wang, and Wenbo Wang* 921

Taxonomies, Applications, and Trends of Mobile Games / *Eui Jun Jeong and Dan J. Kim* 928

Technology Intervention Perspective of Mobile Marketing, A / *Dennis Lee and Ralf Muhlberger* 933

3G Commercial Deployment / *Mugen Peng, Shuping Chen, and Wenbo Wang* 940

Transaction Management in Mobile Databases / *Ziyad Tariq Abdul-Mehdi, Ali Bin Mamat, Hamidah Ibrahim, and Mustafa M. Dirs* 947

Ubiquitous and Pervasive Application Design / *M. Bakhouya and J. Gaber* 954

"Umbrella" Distributed-Hash Table Protocol for Content Distribution, The / *Athanasios-Dimitrios Sotiriou and Panagiotis Kalliaras* 960

Understanding Multi-Layer Mobility / *Sasu Tarkoma and Jouni Korhonen* 966

Using Mobile Devices for Electronic Commerce / *Raul Fernandes Herbster, Hyggo Almeida, and Angelo Perkusich* 974

Using Mobile Devices to Manage Traffic Infractions / *Stefânia Marques, Sabrina Souto, Miguel Queiroga, Hyggo Almeida, and Angelo Perkusich* 978

Using Service Proxies for Content Provisioning / *Panagiotis Kalliaras and Anthanasios-Dimitrios Sotiriou* 981

Verifying Mobile Agent Design Patterns with RPOO / *Elthon Allex da Silva Oliveira, Emerson Ferreira de Araújo Lima, and Jorge César Abrantes de Figueiredo* 987

Virtualization and Mobility in Client and Server Environments / *Eduardo Correia* 996

Voice Recognition Intelligent Agents Technology / *Călin Gurău* 999

Wi-INET Model for Achieving M-Health Success, The / *Nilmini Wickramasinghe and Steve Goldberg* 1004

Wireless Access Control System Using Bluetooth / *Juliano Rodrigues Fernandes de Oliveira, Rodrigo Nóbrega Rocha Xavier, Yuri de Carvalho Gomes, Hyggo Almeida, and Angelo Perkusich* 1011

Wireless Client Server Application Model Using Limited Key Generation Technique / *Rohit Singh, Dhilak Damodaran, and Phu Dung Le* 1015

Wireless Network Security / *Kevin Curran and Elaine Smyth* 1022

Wireless Security / *Meletis Belsis, Alkis Simitsis, and Stefanos Gritzalis* 1028

Wireless Sensor Networks / *Antonio G. Ruzzelli, Richard Tynan, Michael O'Grady, and Gregory O'Hare* 1034

Wireless Technologies for Mobile Computing and Commerce / *David Wright* 1038

Workflow Management Systems in MANETs / *Fabio De Rosa, Massimiliano de Leoni, and Massimo Mecella* 1043

XML-Based Languages for Multimodality in Mobile Environments / *Danilo Avola, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni* 1050

Contents
by Topic

3G

Interworking Architectures of 3G and WLAN / *Ilias Politis, Tasos Dagiuklas, Michail Tsagkaropoulos, and Stavros Kotsopoulos* 357

Protocol Analysis for the 3G IP Multimedia Subsystem / *Muhammad Alam* 778

Protocol Replacement Proxy for 2.5 and 3G Mobile Internet / *Victor Khashchanski, Andrei Kustov, and Jia Lang* 785

Three 3G Commercial Deployment / *Mugen Peng, Shuping Chen, and Wenbo Wang* 940

Adhoc Network

Data Caching in Mobile Ad-Hoc Networks / *Narottam Chand, R.C. Joshi, and Manoj Misra* 172

Mobile Ad-Hoc Networks / *Moh Lim Sim, Choong Ming Chin, and Chor Min Tan* 424

Workflow Management Systems in MANETs / *Fabio De Rosa, Massimiliano de Leoni, and Massimo Mecella* 1043

Converging Technology

Acoustic Data Communication with Mobile Devices / *Victor I. Khachtchanski and Andrei Kustov* 15

Applications Suitability on PvC Environments / *Andres Pablo Flores and Macario Polo Usaola* 57

Bio-inspired Approach for Cellular Systems, A / *Mostafa El-Said* 63

Convergence Technology for Enabling Technologies / *G. Sivaradje, I. Saravanan, and P. Dananjayan* 149

Decision Analysis for Business to Adopt RFID / *Koong Lin, Chad Lin, and Huei Leu* 178

Distributed Web GIS / *Jihong Guan, Shuigeng Zhou, Jiaogen Zhou, and Fubao Zhu* 213

Enabling Technologies for Pervasive Computing / *João H. Kleinschmidt and Walter da Cunha Borelli* 272

Man-Machine Interface with Applications in Mobile Robotic Systems / *Milan Kvasnica* 403

Mobile Users in Smart Spaces / *Loreno Oliveira, Hyggo Almeida, and Angelo Perkusich* 621

Modular Sensory System for Robotics and Human-Machine Interaction / *Milan Kvasnica* 651

Mutual Biometric Authentication / *Mostafa El-Said* 688

Next-Generation Mobile Technologies / *Chor Min Tan, Choong Ming Chin, and Moh Lim Sim* 700

Optimal Timer for Push to Talk Controller, An / *Muhammad Tanvir Alam* 724

Pen-Based Mobile Computing / *Bernie Garrett* 754

Pest Activity Prognosis in the Rice Field / *Nureize Arbaiy, Azizul Azhar Ramli, Zurinah Suradi, and Mustafa Mat Deris* 763

Using Mobile Devices to Manage Traffic Infractions / *Stefânia Daisy Canuto Marques, Sabrina de Figueirêdo Souto, Miguel Queiroga Filho, Hyggo Almeida, and Angelo Perkusich* 978

Human Factor

Academic Activities Based on Personal Networks Deployment / *Vasileios S. Kaldanis, Charalampos Patrikakis, and Vasileios Protonotarios* 1

Adoption and Diffusion of M-Commerce / *Ranjan Kini and Subir Bandyopadhyay* 32

Adoption of M-commerce Devices by Consumers / *Humphry Hung and Vincent Cho* 38

Browser-Less Surfing and Mobile Internet Access / *Gregory J. Fleet and Jeffery G. Reid* 78

Gender Difference in the Motivations of Mobile Internet Usage / *Shintaro Okazaki* 296

M-Commerce Technology Perceptions on Technology Adoptions / *Reychav Iris and Ehud Menipaz* 413

Mobile Commerce Adoption Barriers / *Pruthikrai Mahatanankoon and Juan Garcia* 461

Mobile Enterprise Readiness and Transformation / *Rahul C. Basole and William B Rouse* 481

Mobile ICT / *Dermot McMeel* 516

Mobile Knowledge Management / *Zuopeng Zhang and Sajjad M. Jasimuddin* 520

Mobile Virtual Communities / *Christo El Morr* 632

Location and Context Awareness

Context-Adaptive Mobile Systems / *Christian Kaspar, Thomas Diekman, and Svenja Hagenhoff* 124

Context-Aware Mobile Geographic Information Systems / *Slobodanka Djordjevic – Kajan, Dragan Stojanovic, and Bratislav Predic* 129

Context-Aware Systems / *Chin Chin Wong and Simon Hoh* 138

iPod as a Visitor's Personal Guide / *Keyurkumar Patel and Umesh Patel* 365

Location-Based Multimedia Content Delivery System for Monitoring Purposes / *Athanasios-Dimitrios Sotiriou and Panagiotis Kalliaras* 381

Location-Based Multimedia Services for Tourists / *P. Kalliaras, A. D. Sotiriou, P. Papageorgiou, and S. Zoi* 387

Location-Based Services / *Péter Hegedüs, Mihály Orosz, Gábor Hosszú, and Ferenec Kovács* 393

Monitoring and Tracking Moving Objects in Mobile Environments / *Dragan Stojanovic, Slobodanka Djordjevic-Kajan, Apostolos N. Papadopoulos, and Alexandros Nanopoulos* 660

Notification Services for Mobile Scenarios / *Michael Decker* 711

Ontology-Based Approach for Mobile Agents Context-Awareness, An / *Nejla Amara-Hachmi and Amal El Fallah Seghrouchni* 717

Partial Global Indexing for Location-Dependent Query Processing / *James Jayaputera* 739

Positioning Techologies for Mobile Computing / *Michael O'Grady and Gregory O'Hare* 769

Privacy Concerns for Indoor Location-based Services / *Leonardo Galicia Jimenez and J. Antonio García-Macías* 773

Providing Location-Based Services under Web Services Framework / *Jihong Guan, Shuigeng Zhou, Jiaogen Zhou, and Fubao Zhu* 789

Semantic Enrichment of Location-Based Services / *Vassileios Tsetsos, Christos Anagnostopoulos, and Stathes Hadjiefthymiades* 856

Sensor Data Fusion for Location Awareness / *Odysseas Sekkas, Stathes Hadjiefthymiades, and Evangelos Zervas* 863

M-Business and M-Commerce

Addressing the Credibility of Mobile Applications / *Pankaj Kartham* 25

Advanced Resource Discovery Protocol for Semantic-Enabled M-Commerce / *Michele Ruta, Tommaso Di Noia, Eugenio Di Sciascio, Giacomo Piscitelli, Francesco Maria Donini* 43

Business and Technology Issues in Wireless Networking / *David Wright* 90

Business Strategies for Mobile Marketing / *Indranil Bose and Chen Xi* 96

Contractual Obligations between Mobile Service Providers and Users / *Robert Willis, Alexander Serenko, and Ofir Turel* 143

Dynamic Pricing Based on Net Cost for Mobile Content Services / *Nopparat Srikhutkhao and Sukumal Kitisin* 220

Factors Affecting Mobile Commerce and Level of Involvement / *Frederick Hong Kit Yim, King-Yin Wong, and Alan Ching-bin Tse* 283

Infrastructural Perspective on U-Commerce, An / *Stephen Keegan, Caroline Byrne, Peter O'Hare, and Gregory O'Hare* 310

Intelligent User Preference Detection for Product Brokering / *Sheng-Uei Guan* 334

Interactive Product Catalog for M-Commerce / *Sheng-Uei Guan and Yuan Sherng Tay* 345

M-Advertising / *Michael Decker* 398

Mobile Agent Protection for M-Commerce / *Sheng-Uei Guan* 429

Mobile Business Applications / *Cheon-Pyo Lee* 442

Mobile Commerce / *Jia Jia Wang and Pouwan Lei* 455

Mobile Computing and Commerce / *Stephanie Teufel, Patrick S. Merten, and Martin Steinert* 466

Mobile E-Commerce as a Strategic Imperative for New Economy / *Mahesh S. Raisinghani* 472

Mobile Payment and the Charging of Mobile Services / *Key Pousttchi and Dietmar Georg Wiedemann* 547

Mobile-Based Advertising in Japan / *Shintaro Okazaki* 635

Mobile-Based Research Methods / *Shintaro Okazaki, Akihisa Katsukura, and Mamoru Nishiyama* 639

NFC-Capable Mobile Devices for Mobile Payment Services / *Stamatis Karnouskos* 706

RFID and Wireless Personal Area Networks for Supply Chain Management / *David Wright* 816

Secure Agent Data Protection for E-Commerce Applications / *Sheng-Uei Guan* 826

Snapshot Assessment of Asia Pacific BWA Business Scenario / *Chin Chin Wong, Chor Min Tan, and Pang Leang Hiew* 906

Technology Intervention Perspective of Mobile Marketing, A / *Dennis Lee and Ralf Muhlberger* 933

Using Mobile Devices for Electronic Commerce / *Raul Fernandes Herbster, Hyggo Almeida, and Angelo Perkusich* 974

Wireless Technologies for Mobile Computing and Commerce / *David Wright* 1038

M-Entertainment

Mobile Hunters / *Jörg Lonthoff* 510

Short Message Service (SMS) as an Advertising Medium / *Shintaro Okazaki* 885

Mobile Phone Gambling / *Mark Griffiths* 553

Multimedia Contents for Mobile Entertainment / *Hong Yan, Laura Wang, and Yang Ye* 669

Mobile Entertainment / *Chin Chin Wong and Pang Leang Hiew* 487

Game-Based Methodology for Collaborative Mobile Applications, A / *Michael Massimi, Craig Ganoe, and John M. Carroll* 291

Mobile Television / *Frank Hartung, Markus Kampmann, Uwe Horn, and Jan Kritzner* 611

Mobile Gaming / *Krassie Petrova* 497

Definitions, Key Characteristics, and Generations of Mobile Games / *Eui Jun Jeong and Dan J. Kim* 185

Taxonomies, Applications, and Trends of Mobile Games / *Eui Jun Jeong and Dan J. Kim* 928

M-Health

Mobile Medical Image Viewing Using 3G Wireless Network / *Carrison KS Tong and Eric TT Wong* 533

Mobile Healthcare Communication Infrastructure Networks / *Phillip Olla* 504

Wi-INET Model for Achieving M-Health Success, The / *Nilmini Wichramasinghe and Steve Goldberg* 1004

M-Learning

Integrating Pedagogy, Infrastructure, and Tools for Mobile Learning / *David M. Kennedy and Doug Vogel* 317

Interactive Wireless Morse Code Learning System, An / *Cheng-Hong Yang, Li Yeh Chuang, Cheng-Huei Yang, and Jun-Yang Chang* 352

M-Learning with Mobile Phones / *Simon So* 419

Mobile Learning / *David Parsons* 525

Mobile Learning Environments / *Paul Crowther and Martin Beer* 528

SMS-Based Mobile Learning / *Krassie Petrova* 899

Mobile Multimedia

Adaptive Transmission of Multimedia Data over UMTS / *Antonios Alexiou, Dimitrios Antonellis, and Christos J. Bouras* 20

Enabling Multimedia Applications in Memory-Limited Mobile Devices / *Raul Fernandes Herbster, Hyggo Almeida, Angelo Perkusich, and Marcos Morais* 260

Enabling Technologies for Mobile Multimedia / *Kevin Curran* 265

Interactive Multimedia File Sharing Using Bluetooth / *Danilo Freire de Santos, José Luís do Nascimento, Hyggo Almeida, and Angelo Perkusich* 341

Mobile Serverless Video Communication / *Hans L. Cycon, Thomas C. Schmidt, and Matthias Wählisch* 589

Mobile Sports Video with Total Users Control / *Dian Tjondronegoro* 596

Mobile Video Transcoding Approaches and Challenges / *Ashraf M. A. Ahmad* 627

Multi-User OFDM in Mobile Multimedia Network / *Ibrahim Al Kattan and Habeebur Rahman Maricar* 682

Perceived Quality Evaluation for Multimedia Services / *H. Kourmaras, E. Pallis, G. Xilouris, A. Kourtis, and D. Martakos* 758

Provisioning of Multimedia Applications across Heterogeneous All-IP Networks / *Michail Tsagkaropoulos, Ilias Politis, Tasos Dagiuklas, and Stavros Kotsopoulos* 796

Radio Resource Management in Convergence Technologies / *G. Sivaradje, I. Saravanan, and P. Dananjayan* 810

Shot Boundary Detection Techniques for Video Sequences / *H. Kourmaras, G. Xilouris, E. Pallis, G. Gardikis, and A. Kourtis* 889

Mobile Phone

Enabling Mobile Chat Using Bluetooth / *Ádrian Lívio Vasconcelos Guedes, Jerônimo Silva Rocha, Hyggo Almeida, and Angelo Perkusich* 249

Keyword-Based Language for Mobile Phones Query Services / *Ziyad Tariq Abdul-Mehdi, and Hussein M. Aziz Basi* 369

Mobile Phone Privacy Issues / *Călin Gurău* 557

Mobile Phone Texting in Hong Kong / *Adams Bodomo* 562

Mobile Phones for People with Disabilities / *Hend Al-Khalifa and AbdulMalik S. Al-Salman* 569

Mobile Telephony in Sub-Saharan Africa / *P. Ifinedo* 605

Mobile Text Messaging Interface for Persons with Physical Disabilities / *Cheng-Hong Yan* 616

Multilingual SMS / *Mohammad Shirali-Shahreza* 666

Smartphone Acceptance among Sales Drivers / *Jengchung V. Chen* 894

Voice Recognition Intelligent Agents Technology / *Călin Gurău* 999

Mobile Software Engineering

Accessibility of Mobile Applications / *Pankaj Kartham* 9

Brain Computer Interfacing / *Diego Liberati* 68

Cache Invalidation in a Mobile Environment / *Say Ying Lim* 102

Content Personalization for Mobile Interfaces / *Spiridoula Koukia, Maria Rigou, and Spiros Sirmakessis* 116

Content Transformation Techniques / *Ioannis Antonellis, Vassilis Poulopoulos, and Christos Bouras* 119

Cooperative Caching in a Mobile Environment / *Say Ying Lim* 154

CORBA on Mobile Devices / *Markus Aleksy, Axel Korthaus, and Martin Schader* 160

Design Methodology for Mobile Information Systems / *Zakaria Maamar and Qusay H. Mahmous* 190

Distributed Heterogeneous Tracking for Augmented Reality / *Mihran Tuceryan and Rajeev Raje* 207

Efficient Replication Management Techniques for Mobile Databases / *Ziyad Tariq Abdul-Mehdi, Ali Bin Mamat, Hamidah Ibrahim, and Mustafa Mat Dirs* 233

Extreme Programming for Mobile Applications / *Pankaj Kartham* 277

Handheld Computing and J2ME for Internet-Enabled Mobile Handheld Devices / *Wen-Chen Hu, Jyh-haw Yeh, I-Lung Kao, and Yapin Zhong* 302

Knowledge Representation in Semantic Mobile Applications / *Pankaj Kamthan* 375

Mobile Agent-Based Discovery System / *Rajeev R. Raje, Jayasree Gandhamaneni, Andrew Olson, and Barrett Bryant* 436

Mobile Processes and Mobile Channels / *Kevin Chalmers, Imed Romdhani, and Jon Kerridge* 576

Mobility and Multimodal User Interfaces / *Christopher J. Pavlovski and Stella Mitchell* 644

Multimodality in Mobile Applications and Services / *Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni* 675

New Transaction Management Model / *Ziyad Tariq Abdul Mehdi, Ali Bin Mamat, Hamidah Ibrahim, and Mustafa Mat Dirs* 693

Patterns for Mobile Applications / *Markus Aleksy and Martin Schader* 744

Semantic Caching in a Mobile Environment / *Say Ying Lim* 849

Software Platforms for Mobile Programming / *Khoo Wei Ju and K. Daniel Wong* 912

Transaction Management in Mobile Databases / *Ziyad Tariq Abdul-Mehdi, Hamidah Ibrahim, Mustafa Mat Dirs, and Ali Bin Mamat* 947

Ubiquitous and Pervasive Application Design / *Mohamed Bakhouya and J. Gaber* 954

Umbrella Distributed Hash Table Protocol for Content Distribution, The / *Athanasios-Dimitrios Sotiriou, and Panagiotis Kalliaras* 960

Understanding Multi-Layer Mobility / *Jouni Korhonen and Sasu A .O. Tarkoma* 966

Verifying Mobile Agent Design Patterns with RPOO / *Elthon Allex da Silva Oliveira, Emerson Ferreira de Araújo Lima, and Jorge C. A. de Figueiredo* 987

Virtualization and Mobility in Client and Server Environments / *Eduardo Correia* 996

XML-Based Languages for Multimodality in Mobile Environments / *Danilo Avola, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni* 1050

P2P

Building Web Services in P2P Networks / *Shuigeng Zhou, Jiaogen Zhou, and Jihong Guan* 84

Mobile File-sharing Over P2P Networks / *Lu Yan* 492

P2P Models and Complexity MANETs / *Boon Chong Seet, Chiew-Tong Lau, and Wen-Jing Hsu* 734

Peer-to-Peer Cooperative Caching in Mobile Environments / *Chi-Yin Chow, Hong Va Leong, and Alvin T.S. Chan* 749

Security

Mobile Public Key Infrastructures / *Ioannis Chochliouros, George K. Lalopoulos, Stergios P. Chochliouros, and Anastasia S. Spiliopoulou* 581

Security Architectures of Mobile Computing / *Kaj Grahn, Göran Pulkkis, Jonny Karlsson, and Dai Tran* 839

Wireless Network Security / *Kevin Curran and Elaine Smyth* 1022

Wireless Security / *Meletis Belsis, Alkis Simitsis, and Stefanos Gritzalis* 1028

Sensor Network

Wireless Sensor Networks / *Antonio Ruzzelli, Richard Tynan, Michael O'Grady, and Gregory O'Hare* 1034

Service Computing

Bridging Together Mobile and Service Oriented Computing / *Loreno Oliveira, Emerson Loureiro, Hyggo Almeida, and Angelo Perkusich* 71

Communicating Recommendations in a Service Oriented Environment / *Omar Khadeer Hussain, Elizabeth Chang, Farookh Khadeer Hussain, and Tharam S. Dillon* 108

Embedded Agents for Mobile Services / *John F. Bradley, Conor Muldoon, Gregory O'Hare, and Michael O'Grady* 243

Service Delivery Platforms in Mobile Convergence / *Christopher Pavlovski and Laurence Plant* 870

Service Provision for Pervasive Computing Environments / *Emerson Loureiro, Frederico Bublitz, Loreno Oliveira, Nadia Barbosa, Hyggo Almeida, Glauber Ferreira, and Angelo Perkusich* 877

Using Service Proxies for Content Provisioning / *P. Kalliaras and A. D. Sotiriou* 981

Wireless Networking

Anycast-Based Mobility / *István Dudás, László Bokor, and Sándor Imre* 51

Cross-Layer RRM in Wireless Data Networks / *Amoakoh Gyasi-Agyei* 165

Distributed Approach for QoS Guarantee to Wireless Multimedia / *Kumar S Chetan, P. Venkataram, and Ranapratap Sircar* 195

Distributed Computing in Wireless Sensor Networks / *Hong Huang* 202

Efficient and Scalable Group Key Management in Wireless Networks / *Yiling Wang and Phu Dung Le* 227

Enabling Mobility in IPv6 Networks / *K. Daniel Wong and Saaidal Razalli Bin Azzuhri* 253

Intelligent Medium Access Control Protocol for WSN / *Haroon Malik, Elhadi Shakshuki, and Mieso Denko* 328

Mobile Cellular Traffic with the Effect of Outage Channels / *Hussein M. Aziz Basi and M. B. Ramamurthy* 446

Mobile Multicast / *Thomas C. Schmidt and Matthias Wählisch* 541

Optimal Utilisation of Future Wireless Resources / *Choong Ming Chin, Chor Min Tan, and Moh Lim Sim* 729

QoS Routing Framework on Bluetooth Networking, A / *Chao Liu, Bo Huang, and Takaaki Baba* 804

Scatternet Structure for Improving Routing and Communication Performance / *Bo Huang, Chao Liu, and Takaaki Baba* 820

Secure Group Communications in Wireless Networks / *Yiling Wang and Phu Dung Le* 832

Standard-Based Wireless Mesh Networks / *Mugen Peng, Yingjie Wang, and Wenbo Wang* 921

Wireless Access Control System Using Bluetooth / *Juliano Rodrigues Fernandes de Oliveira, Rodrigo Nóbrega Rocha Xavier, Luiz Paulo de Assis Barbosa, Yuri de Carvalho Gomes, Hyggo Almeida, and Angelo Perkusich* 1011

Wireless Client Server Application Model Using Limited Key Generation Technique / *Rohit Singh, Dhilak Domodaran, and Phu Dung Le* 1015

Foreword

Let us borrow this quote from the British humorist and cartoonist Ashleigh Brilliant to summarize the role of mobility in the development of the information society: "Unless you move, the place where you are is the place where you will always be." In more serious terms, it is fundamental to recognize that today's economic and societal progress is primarily dependent on the technological ability to sustain and facilitate the mobility of persons, physical goods (let us not forget, for instance, that the probably most critical component of global commerce today is deep sea shipping) and digital information (data and programs).

Recent years have witnessed a rapid growth of interest in mobile computing and communications. Indicators are the rapidly increasing penetration of the cellular phone market in Europe, and the mobile computing market is growing nearly twice as fast as the desktop market. In addition, technological advancements have significantly enhanced the usability of mobile communication and computer devices. From the first CT1 cordless telephones to today's Iridium mobile phones and laptops/PDAs with wireless Internet connection, mobile tools and utilities have made the life of many people at work and at home much easier and more comfortable. As a result, mobility and wireless connectivity are expected to play a dominant role in the future in all branches of economy. This is also motivated by the large number of potential users (a U.S. study reports of one in six workers spending at least 20 percent of their time away from their primary workplace, similar trends are observed in Europe). The addition of mobility to data communications systems has not only the potential to put the vision of "being always on" into practice;- but has also enabled new generation of services, for example, location-based services.

Mobile commerce leveraging the mobile Web and mobile multimedia is precisely the ability to deploy and utilize modern technologies for the design, development and deployment of a content rich, user and business friendly, integrated network of autonomous, mobile agents (here "agent" is to be taken in the sense of persons, goods and digital information).

I am delighted to write the foreword to this encyclopedia, as its scope, content and coverage provides a descriptive, analytical, and comprehensive assessment of factors, trends, and issues in the ever-changing field of mobile computing and commerce. This authoritative research-based publication also offers in-depth explanations of mobile solutions and their specific applications areas, as well as an overview of the future outlook for mobile computing.

I am pleased to be able to recommend this timely reference source to readers, be they researchers looking for future directions to pursue when examining issues in the field, or practitioners interested in applying pioneering concepts in practical situations and looking for the perfect tool.

Ismail Khalil Ibrahim,
Johannes Kepler University Linz, Austria
January 2007

Preface

Nowadays, mobile communication, mobile devices, and mobile computing are widely available. Everywhere people are carrying mobile devices, such as mobile phones. The availability of mobile communication networks has made a huge impact to various applications, including commerce. Consequently, there is a strong relationship between mobile computing and commerce. The *Encyclopedia of Mobile Computing and Commerce* brings to readers articles covering a wide range of mobile technologies and their applications.

Mobile commerce (m-commerce) is expanding, and consequently the impact to the overall economy is considerable. However, there are still many issues and challenges to be addressed, such as mobile marketing, mobile advertising, mobile payment, mobile authorization using voice, and so on. Providing users with more intelligent product catalogues for browsing on mobile devices and product brokering also plays an important role in m-commerce. Furthermore, the impact mobile devices give to the supply chain must be carefully considered. This includes the use of emerging mobile technology, such as RFID, sensor network, and so forth.

A wide range of mobile technology is available for m-commerce. Mobile phones are an obvious choice. Additionally, there are many different kinds of mobile phones sold in the market, some of which are labelled as smartphones. There is much research conducted in conjunction with the use of mobile phones. Mobile phone text messaging and SMS are common among mobile users. Subsequently, the use of text messaging and SMS enriches m-commerce, including the ability to support multilingual text messaging. Mobile phone supporting disability has also been a focus lately, which focuses on text messaging to disabled people. More advanced applications now require additional services, such as chatting using Bluetooth, mobile querying, and voice recognition. Mobile privacy issues are also still an important topic.

Apart from mobile phones, there is a wide variety of mobile technology, some of which are mobile robots, RFID, pen-based mobile computing, and so forth. Many advanced applications have been developed utilizing these technologies. Current research has been focusing on man-machine interfaces and sensory systems, particularly for mobile robots, biometric and voice based authentication, traffic infractions, and so forth. The context of smart spaces also gives a new dimension to mobile technology.

The use of mobile technology in entertainment is growing rapidly. Some examples include mobile phone gambling, mobile collaborative games, mobile television, mobile sport videos, and mobile hunting incorporating location-based information. The list is expanding as the technology is advancing. Understanding the success factors for mobile gaming and other entertainment is equally important as the technical aspects of the technology itself.

Videos and multimedia undoubted play an important role in mobile entertainment. Video technologies, such as mobile video sequencing, mobile video transcoding, and mobile video communications, have been studied extensively. One of the main limitations of mobile devices is the limited memory capacity, which has to be carefully addressed, especially in the context of mobile multimedia, because these kinds of applications generally require large amount of spaces. Beside videos, radio technology should not be neglected either.

There are many other applications of mobile technology. For example, the use of mobile technology in health, called m-health, is expanding. Mobile medical imaging is made possible thru the use of 3G wireless network. Another example is the use of mobile technology in learning, called m-learning, such as the use of SMS and text messaging, although some still argue whether m-learning is the way to go in learning, while others are still looking at how to combine the infrastructures and tools with pedagogy.

Developing mobile applications requires a novel software engineering approach. The design for mobile information systems is still maturing. Some researchers are still formulating design patterns for mobile applications, while others are focusing on the user interface aspects. Programming for handheld devices is quite common to use various programming languages and tools, including Java micro edition, J2ME, Corba, and Extreme programming. Since the device generally has a small screen, content transformation and content personalization need to be examined. Other forms of interfaces, including brain computing interfacing, are also interesting. Mobile databases and XML-based mobile technology have received some degree of attention as well.

Other issues that have been incorporated into mobile technologies include mobile agents, service-oriented computing, and various forms of caching, such as peer-to-peer, cooperative, and semantic caching. Service delivery and resource discovery are gaining their popularities too. Security—especially in a mobile environment—should not be neglected. Some work on mobile PKI and limited key generations has been carried out by a number of researchers in order to contribute to advancing m-commerce.

The impact of mobile technology in commerce needs to be evaluated, including its socio-psychological influence and technological adoption and diffusion, as well as readiness and transformation. We need to understand the adoption, barrier, and influencing factors of m-commerce. Some gender issues have been pointed out by some researchers.

All of the abovementioned applications will not be made possible without addressing the advancement of mobile networks. Most of the articles in this encyclopedia may be categorized into the mobile network and communication category. 3G architectures have made their entries lately. Mobile ad-hoc network, IPv6 and P2P are also maturing. Some new work in wireless sensor network is presented.

Last but not least, mobile technology and its applications will not be complete without mentioning location-aware and context-aware. New technologies in positioning; either indoor or outdoor, as well as tracking of moving objects, are presented. Some applications of location-aware include ad-hoc mobile querying, use of iPod as a tourist guide, location-based multimedia for monitoring purposes, and location-based multimedia for tourists. Some notable context-aware applications are notification services, context-aware mobile GIS, and semantic mobile agents for context-aware applications.

As a final note, the *Encyclopedia of Mobile Computing and Commerce* covers a broad range of aspects pertaining to mobile computing, mobile communication, mobile devices, and various mobile applications. These technologies and applications will shape mobile computing and commerce into a new era of the 21st century whereby mobile devices are not only pervasive and ubiquitous, but also widely accepted as the main tool in commerce.

David Taniar
Melbourne, Australia
January 2007

Acknowledgments

I would like to acknowledge the help of all involved in the collation and review process of the encyclopedia, without whose support the project could not have been satisfactorily completed.

I would like to thank all the staff at IGI, whose contributions throughout the whole process, from inception of the initial idea to final publication, have been invaluable. In particular, our thanks go to Kristin Roth, who kept the project on schedule by continuously monitoring the progress on every stage of the project, and to Mehdi Khosrow-Pour and Jan Travers, whose enthusiasm initially motivated me to accept their invitations to take on this project. I am also grateful to my employer Monash University for supporting this project.

A special thank goes to Mr. John Goh of Monash University, who assisted me in almost the entire process of the encyclopedia: from collecting and indexing the proposals, distributing chapters for reviews and re-reviews, constantly reminding reviewers and authors, liaising with the publisher, to many other housekeeping duties, which are endless.

I would also like to acknowledge the assistance and advice from the editorial board members. In closing, I wish to thank all of the authors for their insights and excellent contributions to this encyclopedia, in addition to all those who assisted us in the review process.

David Taniar
Melbourne, Australia
January 2007

About the Editor

David Taniar received a PhD degree in computer science from Victoria University, Australia, in 1997. He is now a senior lecturer at Monash University, Australia. He has published more than 100 research articles and co-authored a number of books in the mobile technology series. He is on the editorial board of a number of international journals in the fields of data warehousing and mining, business intelligence and data mining, mobile information systems, mobile multimedia, Web information systems, and Web and grid services.

Academic Activities Based on Personal Networks Deployment

Vasileios S. Kaldanis
NTUA, Greece

Charalampos Z. Patrikakis
NTUA, Greece

Vasileios E. Protonotarios
NTUA, Greece

INTRODUCTION

Personal networking has already become an increasingly important aspect of the unbounded connectivity in heterogeneous networking environments. Particularly, personal networks (PNs) based on mobile ad-hoc networking have seen recently a rapid expansion, due to the evolution of wireless devices supporting different radio technologies. Bluetooth can be considered as the launcher of the self-organizing networking in the absence of fixed infrastructure, forming pico nets or even scatternets. Similar other wireless technologies (e.g., WiFi) attract a lot of attention in the context of mobile ad hoc networks, due to the high bandwidth flexibility and QoS selection ranges they feature, leveraging the path to develop advanced services and applications destined to the end user and beyond. Furthermore, personal networks are expected to provide a prosperous business filed for exploitation to third-party telecom players such as service and content providers, application developers, integrators, and so forth.

In this article, a personal-to-nomadic networking case is presented. Academic PN (AcPN) is a generic case that aims to describe several situations of daily communication activities within a university campus or an extended academic environment through the support of the necessary technological background in terms of communication technologies. The concept is straightforward: a number of mobile users with different characteristics and communication requirements ranging from typical students to instructors and lecturers, researchers and professors, as well as third parties (e.g., visitors, campus staff), are met, work, interact, communicate, educate, and are being educated within such an environment. This implies the presence of a ubiquitous wireless personal networking environment having nomadic characteristics. Several interesting scenarios and use cases are analyzed, along with a number of proposed candidate mobile technology solutions per usage case.

The article is organized as follows: first, a general description of the academic case is presented identifying examples of typical communication activities within an academic environment; the technical requirements necessary for a successful deployment of personal area network (PAN)/PN technologies within the academic environment are also listed. Next, specific deployment scenarios are presented, followed by a business analysis. The article closes with a concluding section.

ACADEMIC CASE DESCRIPTION

The AcPN case describes several situations of daily communication activities, taking place within a typical university campus environment. Members of the academic community, such as students, make use of personal networking concepts and related technologies to acquire and maintain constant connectivity among them or with local or remote networks, and utilize offered services—applications discovered at their point of presence. In this fashion, they may exchange files on the move, interact with each other in different ways (e.g., messaging, audio/videoconference), connect to a home desktop PC to download a missing file, or configure remotely a project installation located in a lab.

The AcPN case aims to support a number of communication activities known in an academic environment. Typical examples of such activities include:

- entering the campus, and making inquiries for local information (maps, buildings, etc.);
- monitoring information updates (announcements, urgent notices, deadlines, events);
- meeting with a colleague/friend/other student mates, exchanging data with others (docs, mp3, video clips, etc.), work management, and so on;
- seeking a friend/colleagues somewhere on campus;
- communicating with a professor/tutor/technical supervisor;
- reporting project results to colleagues and real-time discussion;

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- borrowing/returning a book from/to the local library;
- performing remote home/office network setup (upon returning home);
- monitoring and controlling a lab experiment/project installation; and
- responding to emergency situations within the campus area (fire drill, medical assistance, etc).

The objective of developing the AcPN case is to provide the academic users with an easy way to perform their everyday work as efficiently as possible—in the least time and with the least cost. The academic entity concept-model used here is very general and includes all different types of academics existing in a typical university environment. These are undergraduates/postgraduates/PhD students, tutors/lecturers/professors, research associates, and third-party entities such as visitors and permanent/temporary staff. The campus infrastructure is supposed to support as many communication technologies as possible to the academic entities roaming on campus, in order to provide a variety of services, featuring flexibility in constructing different networking configurations. These technologies could range from short-distance wireless protocols (Bluetooth, infrared) to large-scale networking solutions such as WLAN or GSM/GPRS and 3G/UMTS.

In any case, academic users can benefit from PN concepts such as P-PAN, PAN/PN, W-PAN, and so forth in order to acquire access to other networks or services. Each user is equipped with a number of wireless communicating devices such as mobile phones, PDAs, laptops, headsets, and mobile storage devices, featuring GSM/GPRS/UMTS Bluetooth and WiFi technologies. These devices can detect and interact with each other in various ways, providing new communication capabilities and fields for different networking configurations.

For example, a student is able to form his own personally attached network or private PAN (P-PAN) by interconnecting his wearable short-range devices (e.g., headset, mp3 player, mobile hard disc, PDA) via Bluetooth or infrared protocol. On a larger scale, the user can also connect to a local network of short-range devices (other users' devices or local wireless printer) becoming part of the existing personal area network, and interact with users in his or her close vicinity who belong to the same network. The student may use his or her mobile device as a GSM/GPRS or UMTS terminal to extend his or her current P-PAN and PAN configuration in order to connect to his or her home DSL network to download an important file from the remote desktop PC. In this case, the student establishes a personal network that can be further used for numerous other remote actions. In the same way, the ubiquitous campus network provider can interconnect all PANs within the campus area and form a "personal"-like network: the campus PN.

Similarly, any other academic user can form one or more PNs dependent on the following parameters:

- the number of interconnecting devices,
- the inherent characteristics of used wireless technologies,
- the connection capabilities per technology in terms of bandwidth and QoS, and
- the requirements imposed by each service used on a particular PN.

Finally, administration of the campus PN is a very important issue for the successful management of attached users in terms of resources and security and successful service provision. Different security levels can be used, according to the trust policy followed when a foreign user (e.g., visitor) is accepted locally in a PAN or globally in the campus PN.

PN CONCEPT IN ACADEMIC CASE

PNs in our case comprise potentially all of a person's devices capable to detect and connect each other in the real or virtual vicinity. Connection is performed via any known and applicable wireless access technology (Bluetooth, infrared, WiFi, MAGNET low/high data rate, WLAN/GSM/GPRS/UMTS, and so on). PN establishment requires an extension of the present and locally detected PAN by the person's attached network (set of person's devices) called private PAN. The physical architecture of the networks and devices (for the AcPN case) has already been mentioned, while all interactions among them is illustrated in the Figure 1.

PNs are configured in an ad hoc fashion, establishing any possible peer-to-peer (P2P) connection among users belonging to the same local PAN and other remote PANs or PNs as well, in order to support a person's private and professional applications. Such applications may be installed and executed on a user's personal device, but also on foreign devices in the same way. PNs consist of communicating clusters of personal digital devices, possibly shared with others and connected through different communication technologies remaining reachable and accessible via at least a PAN/PN. Obviously, PANs have a limited geographical coverage, while PNs have unrestricted geographical span, incorporating devices into the personal environment, regardless of their physical or geographical location. In order to extend their access range, they need the support of typical infrastructure-based and ad-hoc mobile networks.

Strict security policies determine PNs' performance. Any visiting (foreign to the local PAN) mobile user bearing his or her own P-PAN may acquire trust and become a member of the locally detected PAN, as long as another member of the same PAN can guarantee his or her proper behavior in

Figure 1. Academic PN concept topology and interactions

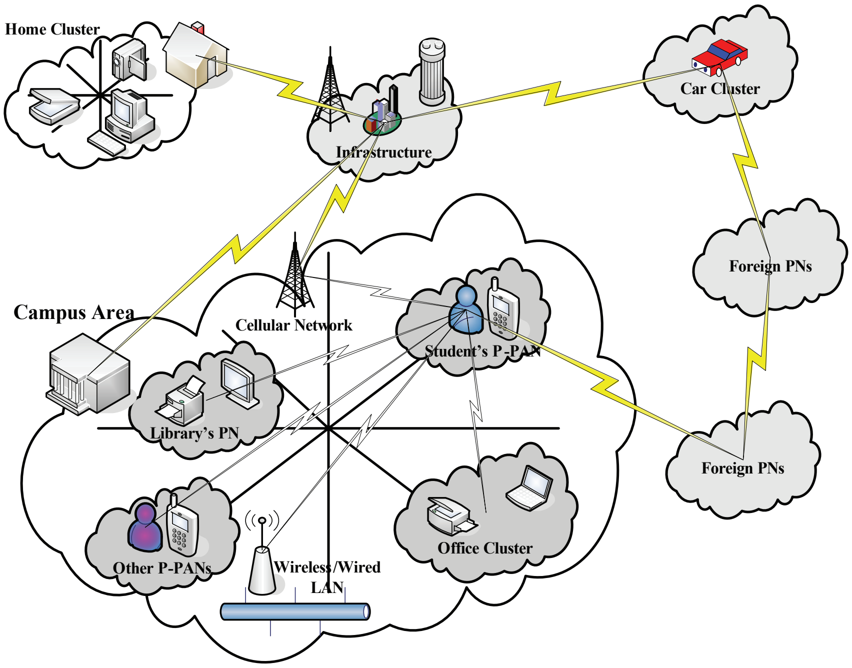

Home Cluster
Infrastructure
Car Cluster
Foreign PNs
Campus Area
Cellular Network
Student's P-PAN
Library's PN
Foreign PNs
Other P-PANs
Wireless/Wired LAN
Office Cluster

Table 1. Devices used in AcPN case

	P-PAN	Home PAN	Office PAN	Campus PN	Laboratory PAN / PN
Desktop PC	-	√	√	√	√
Laptop	Optional	Optional	Optional	Optional	√
PDA	√	√ (Optional)	√ (Optional)	√ (Optional)	√ (Optional)
Mobile Phone	√	Optional	Optional	√	√
MP3 Player	√	√	Optional	Optional	-
Wireless Headset	√	-	-	-	√
Printer	-	√	√	-	√
Scanner	-	√	√	-	√
Mobile Hard Drive	-	√ (Optional)	√ (Optional)	-	√
Camera	-	√	√	-	√
DVD R/Player	-	√	-	-	√
Wall Screen	-	√	√	-	√
Sensor Tx/Rx	-	-	-	-	√

this network. In this way, the new user can become trusted and behave as any other existing user in the PAN. Similar mechanisms exist for AAA functionalities in other clusters and PN domains as well.

A list of important devices for the use cases listed formerly is summarized in Table 1.

- **Desktop PC and Laptop:** High processing power, unlimited power supply, high storage capacity, graphical UI, support of 802.11/Ethernet/Bluetooth, HDR, Internet connectivity (TCP/IP, UDP, etc.), database software, and so forth.
- **PDA:** Low processing power, unlimited power supply, high storage capacity, graphical UI, HDR/LDR,

support 802.11/Ethernet/Bluetooth/56K, support of TCP/IP, security software integrated, low weight, and so forth.
- **High-Featured Mobile Phone Device:** Low processing power, low power consumption, moderate storage space, support of wireless protocols (Bluetooth/GSM/CDMA), IrDA support, cellular connectivity (GSM/GPRS/UMTS), WiFi/WLAN connectivity, portability, synchronization with other devices.
- **Printer:** Support of various wireless technologies (Bluetooth, IrDA, etc), wired networking, and so forth.
- **MP3 Player:** Weak power supply, moderate storage capacity, support of basic wireless access technologies (Bluetooth/WiFi), HDR/LDR, large battery power consumption, low recharging time, handy UI and control, high sound quality, and so forth.
- **Wireless Headset:** Support of basic wireless access technologies (Bluetooth, IrDA), LDR, and so forth.
- **Wireless Sensors:** Low power consumption, wireless interconnectivity (Bluetooth, IrDA, etc.), LDR, large operation life flexible functionality, light weight device, low volume, remotely controllable, and so forth.

It should be noted that currently, there is ongoing work on specifying devices that support new protocols (especially in the wireless physical layer), and the expansion of the use cases to current networking technologies is also still under development.

SCENARIOS AND USE CASES

The scenario generation procedure has been based on the obtained results from an end user workshop held at the NTUA campus. The workshop participants were academic people coming from different knowledge backgrounds and professions (undergraduates/MSc students, PhD candidates/research associates, tutors, lecturers, professors, and visitors). During the workshop all participants had the chance to exchange thoughts and express their own needs regarding communication solutions and services they wish or expect to have within a typical campus area environment.

Login to the Ubiquitous Campus PN Network

This is a fundamental case for the AcPN, since it presents the most important thing an AcPN user must do if he wants to utilize services and applications available in the university domain (single campus or a set of campuses belonging to the same organization).

According to this case, the AcPN user must login to the campus network via his mobile device mainly in two cases: whenever he reaches the real campus area physically (e.g., by car, by bus, or by foot) via a locally detected campus PAN or remotely via a PN which he has previously established dynamically with the campus PN. The AcPN could be a registered user to the campus network (e.g., student, lecturer, researcher, or permanent staff) or a foreign (third-party) user (e.g., visitor) who should follow a registration procedure before attaching to the local network. The login procedure is required for the AcPN case in order to maintain a certain level of security, which is higher for locally connected users in contrast to remotely connected ones. After successful login, the AcPN users can immediately be informed by the campus PN administrator for urgent messages from their colleagues, reminders, scheduled power outages, and other important messages of general importance.

Information Update and Real-Time P2P Interaction

In this use case, an AcPN member, after logging into the campus PN network, wishes to have access to any available local services and applications according to his or her educational activity (e.g., student, researcher). At the same time, he or she can be informed about course announcements, important notices (e.g., deadline extensions, change of lecture classrooms, etc.) from the student office or from any other local online source related to his or her studies. Furthermore, using a mobile device he or she may directly connect to a course database to download important files such as handouts, past papers, presentations, or any other material in electronic form. In P2P fashion, the student may have the chance to see on his or her device who is currently roaming into the campus area from among his contact people (friends, colleagues, tutors, etc.) and to interact with them in various ways. He may also publish hello messages everywhere he wants to, arrange a meeting (physical or not) on the fly, be informed by other people who also "see" him on their devices, exchange files with friends (mp3s, pictures, video clips), send an important file to a colleague or to his or her technical supervisor, setup an audio/video conference, and so forth.

Using a Trusted PAN to Connect to Other Networks

In this scenario, a mobile user who is not a member in the campus PN currently lies within the campus and wishes to get an Internet connection or to acquire access to the local network for several reasons (e.g., utilize local services, get library access, view local events, etc.). This user is considered a foreign user, since he does not belong to the campus PN or to any other local PAN, as privileged campus PN members do. Obviously, the foreign user is considered by the campus network as a third party-user or a visitor and in some way has to be accepted by the campus PN administrator into the ubiquitous local network. This can be done directly or indirectly. In the direct way, the user can be connected to a locally detected PAN at its point of presence if another registered user of the same PAN can guarantee his or her proper and safe behavior. In other words, the foreign user may be attached to any PAN and consequently to the campus PN if another user of the same PAN can verify him or her as a trusted entity and provide him or her with access rights characterized by the basic required security level. In case the foreign user violates the invitation policy agreement, he or she may be warned or even banned by any other PAN user reporting the event to the campus PN administrator. Then the user who signed his or her trustworthiness may lose credits on his or her membership to the campus PN, or his or her authorization provision to other users in the future may be suspended for some period. Following the indirect way, the foreign user may use the local wide area network (e.g., WLAN) to ask for a temporary registration from the campus PN administrator. For example he or she may use a credit card to register to the ubiquitous campus PN network; buy connection time duration; service access rights, bandwidth, and QoS; and so on. In this way, the registered foreign user may be accepted by any other PAN anywhere in the campus, gaining access to the allowed local services in general. This type of user cannot access individual department resources and services (e.g., engineering department database, ftp software, etc.) but only allowed services for third-party users (e.g., library access, local knowledge base intranets, projects, etc.).

Remote Laboratory Monitoring and Control

This is the case where a remote monitoring and controlling of a procedure taking place in a location is required using PAN/PN technologies. Particularly, a group of scientists (students, researchers, professors, etc.) is performing a lab experiment that is long lasting, and the overall progress and results need to be monitored continuously on a 24-hour basis. Furthermore, it is required that according to the collected ongoing results, some experiment parameters may be changed dynamically (locally or remotely). The scientific group must have continuous communication using their mobile devices independent of their point of presence, in order to discuss the change of parameters whenever needed to do so. In this case we consider that there is no physical presence by any member to the lab location and the procedure runs remotely using PAN/PN.

The experiment consists of a number of wireless sensors attached on the examined sample under test, forming a P-PAN which sends reports to a report collector. The report collector enriches the raw report signals and forwards them to the central processing device (high processing power desktop PC) where the experiment software is running. The central processing device sends formatted reports to a local database for data warehousing purposes, while reporting results to the scientific group using the lab PAN as well. Each member of the scientific group has been attached to the lab PAN forming individual PNs and also maintains a direct online connection with the other members for results discussion. Depending on the results, if a parameter change is decided, the user responsible for the experiment sends the required commands to the command executor device, which runs an external application controlling the interaction functionality with the sample under test. The change is verified and archived wirelessly into the database, again using the lab PAN, while a report is sent back to the group about its successful command execution.

Future Library Loaning and Reservation

This scenario presents a proposed loaning and reservation system for academic libraries in the future. In this case, the reservation and loaning of a book title may be performed based on the well-known Web service (via the library Web site) and the campus PN infrastructure. The campus PN consists of all PAN/PN clusters in different departments (or offices/labs, etc.) or smaller departmental libraries and the on-campus users equipped with mobile devices.

According to this scenario, a requestor for a book is an on-campus entity (normal/MSc/PhD student, research fellow), who is using his or her mobile device and the campus networking infrastructure to get access to the local online library database. The requestor should also be a registered member of the campus PN with a stored profile in the university database already logged in. This profile entry automatically enables a number of useful privileges according to the AcPN user type (user profession) that allows him or her to access specific applications and services.

An example use scenario is the following: a requestor gets informed by the library service on his mobile that a requested book is currently loaned and has been delayed to return (i.e., for a day). He is also notified about the priority in the request queue (if any exists) for that title. After that, the system generates an urgent message and forwards it to the loaner of the book using the campus PN. The system, using a tracing mechanism regarding the user status-location, is aware that the loaner is currently active and able to receive notifications via the campus PN, so it prefers to notify the user in this way. The loaner must provide as soon as possible a new book returning date to the library system if he does not want his membership to be blacklisted or in the worst case banned from the campus PN database. Hence, the loaner provides as the new returning date a specific time during the same day. The system forwards the new returning date to the requestor and provides a validity period for his request. After that period, his request is no longer valid and a next requestor (on the queue) gets the right to reserve that book. When finally the book returns to the library desk, the system via the campus PN notifies the active requestor about the book availability and his validity period to come and collect it. The requestor may provide himself as the collecting person or another registered PN user.

Remote Course Exam Participation and Distant Learning

Finally, using this case, a student currently away from the campus area for several reasons (urgent reasons, recuperation in hospital, etc.) has the option to participate remotely in her course exams using the PN technology. At her current location, she as to scan for a local PAN to attach or to search for another local wireless Internet connection means (e.g., WLAN, UMTS, WiFi). Then she must setup a PN with the campus network, logon to the campus PN using her student account, and connect to the local examinations server who has privately published an exams-related session link for such cases. Then after authorizing and authenticate herself, she must download the support software for this online session or any other auxiliary utility supplied by the exam center administrator, install it properly, and directly connect to the exam server before the actual start time of the exams. It is supposed that she has already applied for a remote exam participation by sending an e-mail to the exam administrator, and on reply she has received all the relevant details—information of that session according to the course requirements (e.g., multiple-choice form), connection bandwidth, QoS, and personal mobile device capabilities (e.g., large viewable display, keyboard, memory, etc). The student using this exam PN session participates remotely in the same way she would if she was present in the real exam center location for the required time period of the exam. It is required that she has an interruptible connection with the campus PN network and particularly with the exam center local server. The student provides her answers to the exam paper questions by ticking the appropriate box in each online XML Web interface, presses the "SEND" button to proceed to the next question, and so on. Each provided answer cannot be changed or undone since it has already been sent to the server and saved to the database system. If any problem occurs (e.g., connection is lost or service application fails), the session state is continuously monitored by the exam administrator and resumed when the problem is solved. At the end, the session is closed and a message informs the student that the application has already completed successfully. The service will later inform the student of her achieved results.

In the same way any possible distant learning activity can be supported using similar PN setups and configurations as long as the remote users can create any possible type of PN with the distant network of interest where a relevant service can run reliably.

BUSINESS PROSPECTS

Many players in mobile business may find PN technology to be a prosperous field to extend the market in many dimensions, ranging from high data rate connectivity solutions to advanced services and Web-based applications. The value chain of the mobile market can be dynamically expanded including more than one network and service providers, integrators, service and application developers, or even small-to-medium network operators.

The AcPN case exploits PN concepts in a very efficient way, allowing the use of well-known wireless technologies and common networking configurations of the present and the future to be used and easily applied. Target users are people actively involved in educational activities who present high expectations from communication technologies such as increased bandwidth, connection flexibility (among different technologies), use of a wide range of services and applications, more personalized devices, large mobile storage capability, interoperability, friendly user-device interface, and so forth.

Based on the collected results from the AcPN end user workshop held in Athens, Greece, a number of important requirements have been identified. These requirements have led to several conclusions regarding the new players in the value chain and the business aspects of PAN/PN concepts within the academic environment. The most important conclusions are:

- **Regarding Network Infrastructure:** The network infrastructure should include the normal mobile networks (GSM/GPRS/UMTS), as well as additional networking infrastructure such as WLAN/WiFi on a single or multi-operator environment and the ubiquitous campus PN operator. The campus PN infrastructure must include networking configurations among all campus PANs (different departments, labs, offices) and possibly other PNs (other campuses of the same organization).
- **Regarding Security:** The campus PN operator is responsible for network security in the supported connections of the wired/wireless domain, user login/logout functionality, mobility support within the campus (or campuses of the same university), and other required PAN/PN operations. If the particular university operates more than one campus, then a university PN is required to interconnect the different campus PNs and support the previous on a higher administrative level, securing of course the communication between the PNs.

- **Regarding Service Aggregators:** In this case, the role of service aggregation and provision to the AcPN users is performed primarily by the campus PN operator and partially by third-party service operators who may have agreements with the campus operator. Any service provided to the campus PN is expected to be controlled and maintained by the unique campus PN operator, which plays the twofold role of the service aggregator and the provider. Other services can be provided on the campus by typical mobile operators through the use of voice, e-mail, SMS, or MMS, but PN services and relevant interconnections must be realized via the campus PN network or service operator.
- **Regarding Terminal Equipment:** This requirement takes into account all the different vendors and manufacturers who provide the terminal devices to the end users. The fact that any AcPN user is supposed to be equipped with his or her own P-PAN requires a number of different featured portable devices coming from different vendors to be used. This is feasible as long as the PAN-proposed standards are supported. (It should be noted that for the air interface, the MAGNET LDR/HDR standard has been proposed.)
- **Regarding End Users:** These can be divided into two types. The first one includes all the normal students (undergraduates, postgraduates, etc.) who wish to use typical (low QoS) applications and services (Web browsing, chat, e-mail, voice, SMS, MMS, etc.) within the campus PN at a low cost. The second user type includes any other academic person or third party (visitors, temporary staff) who wish to have (and are willing to pay for) a higher bandwidth wireless connection or access to QoS demanding services such as (real-time) audio/videoconference, streaming applications, and so on. Such users could be professors, tutors, researchers, associates, or general university employees who use telecom technology to communicate with their work contacts for several reasons.

CONCLUSION

The academic case is very promising for the future deployment of PN technologies for many important reasons. First of all, it attempts to combine and reuse efficiently almost any wireless access technologies of the present with proposed ones for the future in many scalable configurations according to the case. Secondly, it provides the option to choose which type of PN could better serve its purposes in terms of connection bandwidth and cost. The user may choose the most efficient way (in terms of cost) to construct his or her own PN; for example, he or she may prefer a relatively cheap WLAN to connect to his or her office rather than a UMTS. Finally, since the use of PN technology might not be possible in some cases without the existence of PAN or P-PAN, the definition of clusters eases the PAN or P-PAN formation as a set of preferable devices, but not all.

ACKNOWLEDGMENTS

The AcPN case was presented, developed, and analyzed in detail within the IST-MAGNET framework (http://www.ist-magnet.org). Specific documents referenced include: MAGNET WP1 Task 1.1, D.1.1.1a, March 2004; MAGNET, WP1 Task 1.4, D1.4.1a, September 2004; MAGNET WP1 Task 1.1, D.1.1.1b, December 2004; MAGNET WP1 Task 1.1, D.1.1.1b, December 2005; and Academic Case Workshop Results, Internal Report D-1.3.1b.

The work acceptance by the academic community is very encouraging and promising for the future. Currently the project group is implementing, based on the previous use cases and scenarios, a number of services.

KEY TERMS

Academic PN (AcPN): Use case descriptive name for a PN exploitation into a typical academic environment.

Cluster: A network of personal devices and nodes located within a limited geographical area (such as a house or a car) which are connected to each other by one or more network technologies and characterized by a common trust relationship between each other.

Context: The information that characterizes a person, place, or object. In that regard, there exist user, environment, and network context. The context information is used to enable context-aware service discovery.

Foreign Device: A device that is not personal and cannot be part of the PN. The device can be either trusted, having an ephemeral trust relationship with another device in the PN, or not trusted at all.

Private Personal Area Network (P-PAN): A dynamic collection of personal nodes and devices around a person.

Personal Area Network (PAN): A network that consists of a set of mobile and wirelessly communicating devices that are geographically close to a person but which may not belong to him.

Personal Device: A device related to a given user or person with a pre-established trust attribute. These devices are typically owned by the user. However, any device exhibiting the trust attribute can be considered as a personal device. The same remarks as those for the personal nodes definition hold for devices.

Personal Network (PN): Network including the P-PAN and a dynamic collection of remote personal nodes and devices in clusters that are connected to each other via interconnecting structures.

Accessibility of Mobile Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The increasing affordability of devices, advantages associated with a device always being handy while not being dependent on its location, and being able to tap into a wealth of information/services has brought a new paradigm to mobile users. Indeed, the *mobile Web* promises the vision of universality: access (virtually) anywhere, at any time, on any device, and to *anybody.*

However, with these vistas comes the realization that the users of the mobile applications and their context vary in many different ways: personal preferences, cognitive/neurological and physiological ability, age, cultural background, and variations in computing environment (device, platform, user agent) deployed. These pose a challenge to the ubiquity of mobile applications and could present obstacles to their proliferation.

This article is organized as follows. We first provide the motivation and background necessary for later discussion. This is followed by introduction of a framework within which accessibility of mobile applications can be systematically addressed and thereby improved. This framework is based on the notions from semiotics and quality engineering, and aims to be practical. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

The issue of accessibility is not new. However, the mobile Web with its potential flexibility on both the client-side and the server-side presents new challenges towards it.

Figure 1 illustrates the dynamics within which the issue of accessibility of a mobile application arises.

We define a *mobile application* as a domain-specific application that provides services and means for interactivity in the mobile Web. For example, education, entertainment, or news syndication are some of the possible domains. The issue of accessibility is intimately related to the user and user context that includes client-side computing environment. To that regard, we define *accessibility* in context of a mobile application as access to the mobile Web by everyone, regardless of their human or environment properties. A *consumer* (user) is a person that uses a mobile application. A *producer* (provider) is a person or an organization that creates a mobile application.

The Consumer Perspective of Mobile Accessibility

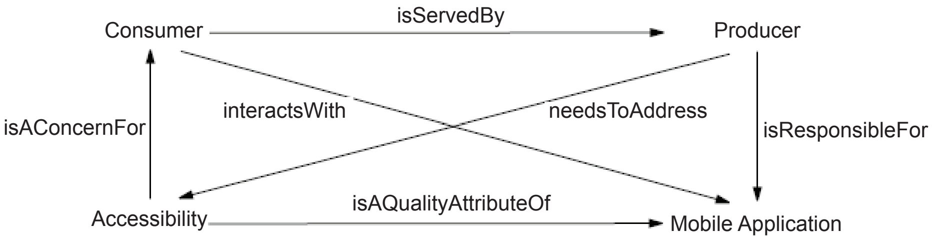
The accessibility concerns of a consumer are of two types, namely human and environment properties, which we now discuss briefly.

Human Properties

Human properties are issues relating to the differences in properties among people. One major class of these properties is related to a person's ability, and often the degree of absence of such properties is termed as a disability. We will use the term "disability" and "impairment" synonymously.

The statistics vary, but according to estimates of the United Nations, about 10% of the world's population is considered disabled. The number of people with some form of disability that do have access to the Internet is in the millions.

Figure 1. The interrelationships between a consumer, a producer, accessibility, and a mobile application



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. A semiotic framework for accessibility of mobile applications

Semiotic Level	Quality Attributes	Means for Accessibility Assurance and Evaluation	Decision Support
Pragmatic	**Accessibility** [T4;E]	• Training in Primary and Secondary Notation • "Expert" Knowledge (Principles, Guidelines, Patterns) • Inspections • Testing • Metrics • Tools	Feasibility
	Comprehensibility, Interoperability, Performance, Readability, Reliability, Robustness [T3;E]		
Semantic	Completeness and Validity [T2;I]		
Syntactic	Correctness (Primary Notation) and Style (Secondary Notation) [T1;I]		

There are several types of disabilities that a producer of a mobile application needs to be concerned with. These can include visual (e.g., low visual acuity, blindness, color blindness), neurological (e.g., epilepsy), auditory (e.g., low hearing functionality, deafness), speech (e.g., difficulties in speaking), physical (e.g., problems using an input device), cognitive (e.g., difficulties of comprehending complex texts and complex structures), cultural/regional (e.g., differences in the use of idioms, metaphors leading to linguistic problems).

Environment Properties

Environment properties are issues relating to different situations in which people find themselves, either temporarily or permanently. These situations could be related to their connectivity, the location they are in, or the device/platform/user agent they are using. For example, a user using a computer in a vehicle shares many of the issues that some people have permanently due to a disability in hand motorics. Or, for example, a user may be accessing the *same* information using a personal digital assistant (PDA) or a cellular phone.

The Producer Perspective of Mobile Accessibility

The motivation for accessibility for a business is to reach as many users as possible and in doing so reduce concerns over customer alienation.

It is the producer of the mobile application that needs to adjust to the user context (and address the issue of accessibility), not the other way around. It is not reasonable for a producer to expect that the consumer environment will be conducive to *anything* that is delivered to him/her. In certain cases, when a consumer has a certain disability, such adaptation is not even possible.

If the success of a mobile application is measured by the access to its services, then improving accessibility is critical for the producers. Still, any steps that are taken by a producer related to a mobile application have associated costs and trade-offs, and the same applies to improvements towards accessibility.

Initiatives for Improving Accessibility in Mobile Contexts

There are currently only a few efforts in systematically addressing accessibility issues pertaining to mobile applications.

There are guidelines available for addressing accessibility (Chisholm, Vanderheiden, & Jacobs, 1999; Ahonen, 2003) in general and language-specific techniques (Chisholm et al., 2000) in particular.

ADDRESSING THE ACCESSIBILITY OF MOBILE APPLICATIONS

To systematically address the accessibility of mobile applications, we take the following steps:

1. View accessibility as a qualitative aspect and address it indirectly via quantitative means.
2. Select a theoretical basis for communication of information (semiotics), and place accessibility in its setting.
3. Address semiotic quality in a practical manner.

Based on this, we propose a framework for accessibility of mobile applications (see Table 1). The external attributes (denoted by E) are extrinsic to the mobile application and are directly the consumer's concern, while internal attributes (denoted by I) are intrinsic to the mobile application and are directly the producer's concern. Since not all attributes corresponding to a semiotic level are on the same echelon, the different tiers are denoted by "Tn."

We now describe each component of the framework in detail.

Semiotic Levels

The first column of Table 1 addresses semiotic levels. Semiotics (Stamper, 1992) is concerned with the use of symbols to convey knowledge.

From a semiotics perspective, a representation such as a mobile resource can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social, each depending on the previous one in that order. The physical and empirical levels are concerned with the physical representation of signs in hardware and communication properties of signs, and are not of direct concern here. The syntactic level is responsible for the formal or structural relations between signs. The semantic level is responsible for the relationship of signs to what they stand for. The pragmatic level is responsible for the relation of signs to interpreters. The social level is responsible for the manifestation of social interaction with respect to signs, and is not of direct concern here.

We note that none of the layers in Table 1 is sufficient in itself for addressing accessibility and intimately depends on other layers. For example, it is readily possible to create a document in XHTML Basic, a markup language for small information appliances such as mobile devices, that is syntactically correct but is semantically non-valid. This, for instance, would be the case when the elements are (mis)used to create certain user-agent-specific presentation effects. Now, even if a mobile resource is syntactically and semantically acceptable, it could be rendered in such a way that it is unreadable (and therefore violates an attribute at the pragmatic level). For example, this could be the case by the use of very small fonts for some text, or the colors chosen for background and text foreground being so close that the characters are hard to discern.

Quality Attributes

The second column of Table 1 draws the relationship between semiotic levels and corresponding quality attributes. We contend that the quality attributes we mention are necessary but make no claim of their sufficiency.

The internal quality attributes for syntactic and semantic levels are inspired by Lindland, Sindre, and Sølvberg (1994). At the semantic level, we are only concerned with the conformance of the mobile application to the domain it represents (that is, semantic correctness or completeness) and at the syntactic level the interest is in conformance with, with respect to the languages used to produce the mobile application (that is, syntactic correctness).

Accessibility belongs to the pragmatic level and depends on the layers beneath it. It in turn depends upon the other external quality attributes, namely comprehensibility, interoperability, performance, readability, reliability, and robustness, which are also at the pragmatic level. Since these are perceived as necessary conditions, violations of any of these lead to a deterioration of accessibility.

Means for Accessibility Assurance and Evaluation

The third column of Table 1 lists the direct and indirect (and not necessarily mutually exclusive) means for assuring and evaluating accessibility:

- **Training in Primary and Secondary Notation:** The knowledge of the *primary notation* of all technologies (languages) is necessary for guaranteeing conformance to tier T1. The Cognitive Dimensions of Notations (CDs) (Green, 1989) are a generic framework for describing the utility of information artifacts by taking the system environment and the user characteristics into consideration. Our main interest here is in the CD of *secondary notation.* This CD is about appropriate use (that is, *style*) of primary notation in order to assist in interpreting semantics. It uses the notions of *redundant recoding* and *escape from formalism* along with spatial layout and perceptual cues to clarify information or to give hints to the stakeholder, both of which aid the tiers T2 and T3. Redundant Recoding is the ability to express information in a representation in more than one way, each of which simplifies different cognitive tasks. It can be introduced in a textual mobile resource by making effective use of orthography, typography, and white space. Escape from Formalism is the ability to intersperse natural language text with formalism. Mobile resources could be complemented via natural language annotations (metadata) to make the intent or decision rationale of the author explicit, or to aid understanding of stakeholders that do not have the necessary technical knowledge. Incidentally, many of the language-specific techniques for accessibility (Chisholm et al., 2000) are in agreement with this CD.
- **"Expert" Body of Knowledge:** The three types of knowledge that we are interested are principles, guidelines, and patterns. Following the basic principles (Ghezzi, Jazayeri, & Mandrioli, 2003; Bertini, Catarci, Kimani, & Dix, 2005) underlying a mobile application enables a provider to improve quality attributes related to tiers T1-T3 of the framework. However, principles tend to be abstract in nature which can lead to multiple interpretations in their use and not mandate conformance. The guidelines encourage the use of conventions and good practice, and could serve as a checklist with respect to which an application could be heuristically or otherwise evaluated. The guidelines

available for addressing accessibility (Chisholm et al., 1999; Ahonen, 2003) when tailored to mobile contexts can be used as means for both assurance and evaluation of accessibility of mobile applications. However, guidelines tend to be more useful for those with an expert knowledge than for a novice to whom they may seem rather general to be of much practical use. The problems in using tools to automatically check for accessibility have been outlined in Abascal, Arrue, Fajardo, Garay, and Tomás (2004). Patterns (Alexander, 1979) are reusable entities of knowledge and experience aggregated by experts over years of "best practices" in solving recurring problems in a domain including mobile applications (Roth, 2002; Van Duyne, Landay, & Hong, 2003). They are relatively more structured compared to guidelines and, if represented adequately, provide better opportunities for sharing and reuse. There is, however, an associated cost of learning and adapting patterns to new contexts.

- **Inspections:** Inspections (Wiegers, 2002) are a rigorous form of auditing based upon peer review that can address quality concerns for tiers T1, T2, and most of T3, and help improve the accessibility of mobile applications. Inspections could, for example, use the guidelines and decide what information is and is not considered "comprehensible" by consumers at-large, or whether the choice of labels in a navigation system enhances or reduces readability. Still, inspections, being a means of static verification, cannot completely assess interoperability, performance, reliability, or robustness. Furthermore, inspections do involve an initial cost overhead from training each participant in the structured review process, and the logistics of checklists, forms, and reports.
- **Testing:** Some form of testing is usually an integral part of most development models of mobile applications (Nguyen, Johnson, & Hackett, 2003). However, due to its very nature, testing addresses quality concerns only at of some of the tiers (T1, subset of T2, subset of T3). Interoperability, performance, reliability, and robustness would intimately depend on testing. Unlike inspections, tool support is imperative for testing. Therefore, testing *complements* but does not replace inspections.
- **Metrics:** In a resource-constrained environment of mobile devices, efficient use of time and space is critical. Metrics (Fenton & Pfleeger, 1997) provide a quantitative means for making qualitative judgments about quality concerns at technical levels. There is currently limited support for metrics for mobile applications in general and for their accessibility (Arrue, Vigo, & Abascal, 2005) in particular. Any dedicated effort of deploying metrics for accessibility measurement would inevitably require tool support, which at present is lacking.
- **Tools:** Tools that have help improve quality concerns at all tiers. For example, tools can help report violations of accessibility guidelines, or find non-conformance to markup or scripting language syntax. However, at times tools cannot address some of the stylistic issues (such as an "optimal" distance between two text fragments that will improve readability) or semantic issues (like semantic correctness of a resource included in a mobile application). Therefore, the use of tools as means for automatic accessibility evaluation should be kept in perspective.

Decision Support

A systematic approach to a mobile application must take a variety of constraints into account: organizational constraints (personnel, infrastructure, schedule, budget, and so on) and forces (market value, competitors, and so on).

A producer would need to, for example, take into consideration the cost of an authoring tool vs. the accessibility support it provides; since complete accessibility testing is virtually impossible, determine a stopping criteria that can be attained within the time constraints before the application is delivered; and so on.

Indeed, the last column of Table 1 acknowledges that with respect to any assurance and/or evaluation, and includes feasibility as an all-encompassing consideration on the layers to make the framework practical. There are well-known techniques such as analytical hierarchy process (AHP) and quality function deployment (QFD) for carrying out feasibility analysis, and further discussion of this aspect is beyond the scope of this article.

FUTURE TRENDS

Much of the development of mobile applications is carried out on the desktop. The tools in the form of software development toolkits (SDK) and simulators such as Nokia Mobile Internet Toolkit, Openwave Phone Simulator, and NetFront Mobile Content Viewer assist in that regard. However, explicit support for accessibility in these tools is currently lacking.

The techniques for accessibility for mobile technologies such as XHTML Basic/XHTML Mobile Profile (markup of information) and CSS Mobile Profile (presentation of information) would be of interest. This is especially an imperative considering that the widely used traditional representation languages such as Compact HTML (cHTML), an initiative of the NTT DoCoMo, and the Wireless Markup Language (WML), an initiative of the Open Mobile Alliance (OMA), have evolved towards XHTML Basic or its extensions such as XHTML Mobile Profile.

Identification of appropriate CDs, and an evaluation of the aforementioned languages for presentation or representation of information in a mobile context with respect to them, would also be of interest.

As mobile applications increase in size and complexity, a systematic approach to developing them arises. Indeed, accessibility needs to be a part of the *entire* lifecycle of a mobile application—that is, in the typical workflows of planning, modeling, requirements, design, implementation, and verification and validation. To that regard, integrating accessibility into "lightweight" process methodologies such as Extreme Programming (XP) (Beck & Andres, 2005) that is adapted for a systematic development of small-to-medium scale mobile applications would be useful. A similar argument can be made for the "heavyweight" case, for example, by instantiating the Unified Process (UP) (Jacobson, Booch, & Rumbaugh, 1999) for medium-to-large scale mobile applications.

Finally, a natural extension of the issue of accessibility is to the next generation of mobile applications, namely mobile applications on the semantic Web (Hendler, Lassila, & Berners-Lee, 2001). The mobile applications for the Semantic Web present unique accessibility issues such as inadequacy of current searching techniques (Church, Smyth, & Keane, 2006) and a promising avenue for potential research.

CONCLUSION

This article takes the view that accessibility is not only a technical concern, it is also a social right. In that context, the issues of credibility and legality are particularly relevant as both are at a higher echelon (social level) than accessibility within the semiotic framework.

Credibility is considered to be synonymous to (and therefore interchangeable with) believability (Hovland, Janis, & Kelley, 1953). Indeed, improvement of accessibility is necessary for a demonstration of *expertise,* which is one of the dimensions (Fogg, 2003) of establishment of credibility of the producer with the consumer.

Accessibility is now a legal requirement for public information systems of governments in Canada, the U.S., Australia, and the European Union. The producers need to be aware of the possibility that, as mobile access becomes pervasive in society, the legal extent could expand to mobile applications.

As is well known in engineering contexts, preventative measures such as addressing the problem *early* are often better than curative measures at late stages when they may just be prohibitively expensive or simply infeasible. If accessibility is to be considered as a first-class concern by the producer, it needs to be more than just an afterthought; it needs to be integral to mobile Web engineering.

REFERENCES

Abascal, J., Arrue, M., Fajardo, I., Garay, N., & Tomás, J. (2004). The use of guidelines to automatically verify Web accessibility. *Universal Access in the Information Society, 3*(1), 71-79.

Ahonen, M. (2003, September 19). Accessibility challenges with mobile lifelong learning tools and related collaboration. *Proceedings of the Workshop on Ubiquitous and Mobile Computing for Educational Communities: Enriching and Enlarging Community Spaces (UMOCEC 2003)*, Amsterdam, The Netherlands.

Alexander, C. (1979). *The timeless way of building.* Oxford, UK: Oxford University Press.

Arrue, M., Vigo, M., & Abascal, J. (2005, July 26). Quantitative metrics for Web accessibility evaluation. *Proceedings of the 1st Workshop on Web Measurement and Metrics (WMM05)*, Sydney, Australia.

Beck, K., & Andres, C. (2005). *Extreme programming explained: Embrace change* (2nd ed.). Boston: Addison-Wesley.

Bertini, E., Catarci, T., Kimani, S., & Dix, A. (2005). A review of standard usability principles in the context of mobile computing. *Studies in Communication Sciences, 1*(5), 111-126.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0.* W3C Recommendation, World Wide Web Consortium (W3C).

Chisholm, W., Vanderheiden, G., & Jacobs, I. (2000). *Techniques for Web content accessibility guidelines 1.0.* W3C Note, World Wide Web Consortium (W3C).

Church, K., Smyth, B., & Keane, M.T. (2006, May 22). Evaluating interfaces for intelligent mobile search. *Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility 2006 (W4A2006)*, Edinburgh, Scotland.

Fogg, B.J. (2003). *Persuasive technology: Using computers to change what we think and do.* San Francisco: Morgan Kaufmann.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The unified software development process.* Boston: Addison-Wesley.

Ghezzi, C., Jazayeri, M., & Mandrioli, D. (2003). *Fundamentals of software engineering* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software metrics: A rigorous & practical approach.* International Thomson Computer Press.

Green, T. R. G. (1989). Cognitive dimensions of notations. In V. A. Sutcliffe & L. Macaulay (Ed.), *People and computers* (pp. 443-360). Cambridge: Cambridge University Press.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The semantic Web. *Scientific American, 284*(5), 34-43.

Hovland, C. I., Janis, I. L., & Kelley, J. J. (1953). *Communication and persuasion.* New Haven, CT: Yale University Press.

Lindland, O. I., Sindre, G., & Sølvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE Software, 11*(2), 42-49.

Nguyen, H. Q., Johnson, R., & Hackett, M. (2003). *Testing applications on the Web: Test planning for mobile and Internet-based systems* (2nd ed.). New York: John Wiley & Sons.

Paavilainen, J. (2002). *Mobile business strategies: Understanding the technologies and opportunities.* Boston: Addison-Wesley.

Van Duyne, D. K., Landay, J., & Hong, J. I. (2003). *The design of sites: Patterns, principles, and processes for crafting a customer-centered Web experience.* Boston: Addison-Wesley.

KEY TERMS

Cognitive Dimensions of Notations: A generic framework for describing the utility of information artifacts by taking the system environment and the user characteristics into consideration.

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Mobile Accessibility: Access to the Web by everyone, regardless of their human or environment properties.

Mobile Resource: A mobile network data object that can be identified by a URI. Such a resource may be available in multiple representations.

Mobile Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering, and management principles, and disciplined and systematic approaches to the successful development, deployment, and maintenance of high-quality mobile Web applications.

Quality: The totality of features and characteristics of a product or a service that bear on its ability to satisfy stated or implied needs.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and their representations.

Acoustic Data Communication with Mobile Devices

Victor I. Khashchanskiy
First Hop Ltd., Finland

Andrei L. Kustov
First Hop Ltd., Finland

INTRODUCTION

One of the applications of m-commerce is mobile authorization, that is, rights distribution to mobile users by sending authorization data (a token) to the mobile devices. For example, a supermarket can distribute personalized discount coupon tokens to its customers via SMS. The token can be a symbol string that the customers will present while paying for the goods at the cash desk. The example can be elaborated further—using location information from the mobile operator, the coupons can only be sent to, for example, those customers who are in close vicinity of the mall on Saturday (this will of course require customers to allow disclosing their location).

In the example above, the token is used through its manual presentation. However, most interesting is the case when the service is released automatically, without a need for a human operator validating the token and releasing a service to the customer; for example, a vending machine at the automatic gas station must work automatically to be commercially viable.

To succeed, this approach requires a convenient and uniform way of delivering authorization information to the point of service—it is obvious that an average user will only have enough patience for very simple operations. And this presents a problem.

There are basically only three available local (i.e., short-range) wireless interfaces (LWI): WLAN, IR, and Bluetooth, which do not cover the whole range of mobile devices. WLAN has not gained popularity yet, while IR is gradually disappearing. Bluetooth is the most frequently used of them, but still it is not available in all phones.

For every particular device it is possible to send a token out using some combination of LWI and presentation technology, but there is no common and easy-to-use combination. This is a threshold for the development of services.

Taking a deeper look at the mobile devices, we can find one more non-standard simplex LWI, which is present in all devices—acoustical, where the transmitter is a phone ringer. Token presentation through acoustic interface along with general solution of token delivery via SIM Toolkit technology (see 3GPP TS, 1999) was presented by Khashchanskiy and Kustov (2001). However, mobile operators have not taken SIM Toolkit into any serious use, and the only alternative way of delivering sound tokens into the phone-ringing tone customization technology was not available for a broad range of devices at the time the aforementioned paper was published.

Quite unexpectedly, recent development of mobile phone technologies gives a chance for sound tokens to become a better solution for the aforementioned problem, compared with other LWI. Namely, it can be stated that every contemporary mobile device supports either remote customization of ringing tones, or MMS, and in the majority of cases, even both, thus facilitating sound token receiving over the air.

Most phone models can playback a received token with only a few button-clicks. Thus, a sound token-based solution meets the set criteria better than any other LWI. Token delivery works the same way for virtually all phones, and token presentation is simple.

In this article we study the sound token solution practical implementation in detail. First, we select optimal modulation, encoding, and recognition algorithm, and we estimate data rate. Then we present results of experimental verification.

ACOUSTIC DATA CHANNEL

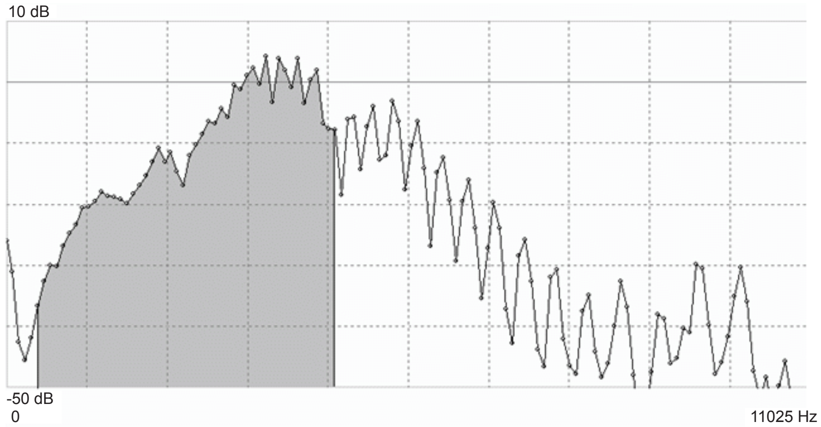
We consider the channel being as follows. The transmitter is a handset ringer; information is encoded as a sequence of sine wave pulses, each with specified frequency and amplitude.

Multimedia message sounds and most ringing tones are delivered as sequences of events in MIDI (musical instrument digital interface) format. A basic pair of MIDI events (note on and note off) defines amplitude, frequency, duration of a note, and the instrument that plays this note. MIDI events can be used to produce information-bearing sound pulses with specified frequency and amplitude.

Widely used support of polyphonic MIDI sequences allows playback of several notes simultaneously. Nonetheless, this has been proved worthless because in order to get

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Frequency response measured with test MIDI sequence in hold-max mode



distinguished, these notes have to belong to different non-overlapping frequency ranges. Then the bit rate that can be achieved would be the same as if wider frequency range was allocated for a single note.

The receiver is a microphone; its analog sound signal is digitized and information is decoded from the digital signal by recognition algorithm, based on fast fourier transform (FFT) technique. FFT is, in our opinion, a reasonable trade-off between efficiency and simplicity.

We investigated acoustics properties of mobile devices. After preliminary comparison of a few mobile phone models, we found that ringer quality is of approximately the same level. All handsets have a high level of harmonic distortions and poor frequency response. The results shown in Figures 1 and 2 are obtained for a mid-class mobile phone SonyEricsson T630 and are close to average.

MIDI-based sound synthesis technology applies limitations on pulse magnitude, frequency, and duration. At the same time, ringer frequency response is not linear and the level of harmonic distortions is very high. Figure 1 shows frequency response measured with a sweeping tone or, to be precise, a tone leaping from one musical note to another. To obtain this, the phone played a MIDI sequence of non-overlapping in-time notes that covered a frequency range from 263 to 4200 Hz (gray area).

The frequency response varies over a 40 dB range, reaching its maximum for frequencies from approximately 2.5 to 4 kHz. Moreover, spectral components stretch up to 11 KHz, which is caused by harmonic distortions. This is illustrated also by Figure 2.

Horizontal axis is time; overall duration of the test sequence is 15 seconds. Vertical axis is sound frequency, which is in range from 0 to 11025 Hz. Brightness is proportional to sound relative spectral density; its dynamic range is 60 dB, from black to white.

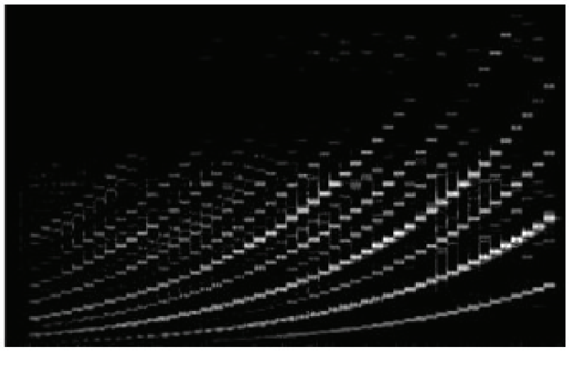
We also found that frequency of the same note may differ in different handsets. Nevertheless, the ratio of note frequencies (musical intervals) remains correct, otherwise melodies would sound wrong.

For a simplex channel with such poor parameters, as reliable a data encoding method as possible is to be used. Frequency shift keying (FSK) is known as the most reliable method which finds its application in channels with poor signal-to-noise ratio (SNR) and non-linear frequency response.

It is not possible to negotiate transfer rate or clock frequency, as it is usually done in modem protocols because acoustic channel is simplex. To make the channel as adaptive as possible, we have chosen to use differential FSK (DFSK), as it requires no predefined clock frequency. Instead, frequency leaps from one pulse to another provide the channel clocking. The difference between frequencies of consecutive pulses determines the encoded value.

Once encoding scheme is selected, let us estimate possible transfer rate before we can find the balance between data

Figure 2. A spectrogram of the test MIDI sequence



transfer rate and channel reliability. Suppose the transmitter generates a sequence of pulses of duration τ, which follow without gaps with repetition frequency f. If each frequency leap between two consequent pulses carries N bits of information, the overall bit rate p is obviously:

$$p = N \bullet f. \tag{1}$$

In DFSK, for each frequency leap to carry N bits, we must be able to choose pulse frequencies from a set of 2^N+1 values. If a pulse frequency can have n values, we will have

$$p = [\log_2(n-1)] \bullet f, \tag{2}$$

where by $[]$ we denote integer part. It follows from (1, 2), that to increase p, we must increase pulse repetition frequency f and the amount of possible values for pulse carrier frequencies n. However, if the recognition is based on spectral analysis, we cannot increase n and f independently. Let us show it. Assume for simplicity that pulse frequency can have any value within frequency range F. Then the number of available values of coding frequencies will be

$$n = [\log_2(F/\Delta f - 1)], \tag{3}$$

where Δf is the minimal shift of pulse frequency between two consecutive pulses. Maximum n is achieved with maximum F and minimum Δf. Both parameters have their own boundaries. Bandwidth is limited by the ringer capabilities, and frequency shift is dependant on pulse repetition frequency f, due to the fundamental rule of spectral analysis (Marple, 1987), which defines frequency resolution δf to be in reverse proportionality to observation time T:

$$\delta f = 1 / T. \tag{4}$$

How can (4) be understood in our case of a sequence of pulses? Having converted the signal into frequency domain, we will get the sequence of spectra. As information is encoded in the frequency pulses, we must determine the pulse frequency for every spectrum. This can only be done with certain accuracy δf called frequency resolution. The longer time T we observe the signal, the better frequency resolution is. So for given pulse duration τ, equation (4) sets the lower limit for frequency difference Δf between two consecutive pulses:

$$\Delta f \geq 1 / \tau = f. \tag{5}$$

This means, that if we increase pulse repetition rate f, then we have to correspondingly increase frequency separation Δf for the consecutive pulses; otherwise the spectral analysis-based recognizing device will not principally be able to detect signal.

Let us now try to estimate the data rate for the system we studied earlier. Figures 1 and 2 show that harmonic distortions are very high, and second and third harmonics often have higher magnitudes than the main tone. Consequently, the coding frequencies must belong to the same octave. Their frequency separation should be no less than defined by (5).

An octave contains 12 semitones, so possible frequency values f_i are defined by the following formula:

$$f_i = f_0 \bullet 2^{i/12},\ i=0...11. \tag{6}$$

The minimum spacing between consecutive notes is for i=1; maximum for i=11.

In our case, we decided to use the fourth octave—as the closest to the peak area of phone ringer frequency response—in order to maximize SNR and thus make recognition easier. For it, f_0= 2093 Hz, and minimum spacing between notes is 125 Hz. Taking the maximum amount of N = 3 (9 coding frequencies), we can estimate transfer rate as:

$$p_{max} = 3 \bullet 125 = 375 \text{ bps}. \tag{7}$$

Recognition Algorithm (Demodulation)

The following algorithm was developed to decode information transferred through audio channel. Analog audio signal from the microphone is digitized with sampling frequency Fs satisfying Nyquist theorem (Marple, 1987). A signal of duration Ts is then represented as a sequence of Ts/Fs samples. FFT is performed on a sliding vector of M signal samples, where M is a power of 2.

- First, sequence of instant power spectra is obtained from the signal using discrete Fourier transform with sliding window (vector) of M samples. To get consecutive spectra overlapped by 50%, the time shift between them was taken $M/2Fs$. Overlapping is needed to eliminate the probability of missing the proper position of a sliding window corresponding to the pulse existence interval, when the pulse duration is not much longer than analysis time significantly (at least twice).
- Second, the synchronization sound is found as sine wave with a constant, but not known in advance frequency, and a certain minimum duration.
- Third, the spectrum composed of maximum values over the spectra sequence (so-called hold-max spectrum) is used to find the pulse carrier frequencies. This step relies on the assumption that used frequency range does not exceed one octave. In other words, the highest frequency is less than twice the value of the lowest one.
- Forth, time cross-sections of spectra sequence at found carrier frequencies are used to recognize moments of sound pulse appearances.

- The last step is reconstruction of encoded bit sequence having the time-ordered set of frequency leaps.

Such an algorithm does not need feedback and can work with unknown carrier frequencies in unknown but limited frequency range. Recognizing the beginning of the transmission is critical for the correct work, so we added "synchronization header" in the beginning of the signal. The length of this header is constant, so the throughput of the system will rise with the message length.

Recognizer Parameters

Here we explain how the parameters of analyzer (***Fs, M***) are defined from that of signal (***f, Δf***). After FFT, we have ***M/2*** of complex samples in frequency domain, corresponding to frequency range from 0 to ***Fs/2***. So for this particular case, frequency resolution obviously equals the difference between the consecutive samples in the frequency domain; namely,

$$df = Fs / M. \tag{7}$$

According to (4), minimum required time of analysis is

$$T = M / Fs. \tag{8}$$

It is obvious, that ***T*** must not exceed burst duration τ. Combining (8) and (5), we get:

$$M / Fs \leq 1 / f \tag{9}$$

On the other hand, frequency resolution ***df*** must not exceed spacing ***Δf*** between carrier frequencies:

$$Fs / M \leq \Delta f \tag{10}$$

Combining (9) and (10), we will finally get:

$$f \leq Fs / M \leq \Delta f \tag{11}$$

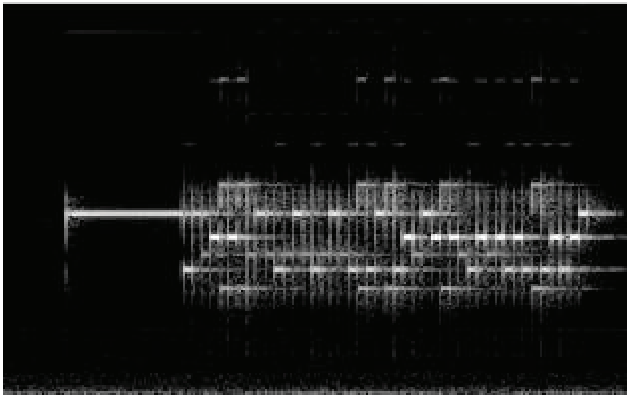
which shows that values of analyzer parameters may be restricted when (5) is close to the equation. This imposes requirements on the sound recognition algorithm to work reliably nearby the "critical points," where the recognition becomes principally impossible.

EXPERIMENTAL RESULTS

We implemented a prototype of acoustic data channel with the mobile phone SonyEricsson T630, whose characteristics are seen in Figures 1 and 2.

For encoding, we developed software that encoded symbol strings in ASCII to melody played by an electric organ. The instrument was chosen from 127 instruments available in MIDI format, because its sound is the closest to the sine wave pulses model we used in calculations. It is maintained at approximately the same level over the whole note duration.

Figure 3. Encoded "hello world"; note the leading synchronization header. Overall duration is approximately 2 seconds.



The recognizer consisted of a Sony ECM-MS907 studio microphone for signal recording, and a conventional PC with a sound card was used for signal analysis. FFT processing was done by our own software.

In the beginning of our experiments, we used the parameters described in the theoretical section. Later we found that at the highest possible transfer rate, data recognition is not reliable. So we gradually increased pulse duration until recognition became reliable. Eventually we selected the following modulation parameters: ***n=5*** (each frequency leap carries two data bits), notes were evenly distributed over the octave (C, D#, F, G, A in musical notation, and they correspond to frequencies 2093, 2489, 2794, 3136, and 3520 Hz), and pulse duration was 46 ms.

Figure 3 shows a spectrogram of recognizable signal from the microphone.

Horizontal axis is time, and overall signal duration is 2 seconds. Vertical axis is frequency, and one can see the leaps between consecutive pulses. Brightness is proportional to the signal intensity.

This example signal carries 88 bits of information (a string "hello world," coded as 11 ASCII characters), which makes the data transfer rate approximately 40 bps. Overhead from the synchronization header is ca. 25%; for longer messages the average transfer rate would be higher.

DISCUSSION

We have managed to implement a reliable data channel from the phone; the advantage of the proposed recognition algorithm is that it can work in the same way for every mobile device, independent on acoustic properties of different brands and models, although encoding frequencies are different.

The channel is principally one way: the handset cannot receive any feedback that can be used, for example, for error correction. Nevertheless, developed recognition algorithm provided good reliability. For a handset placed 30 cm from the microphone, in a room environment, recognition was 100% reliable. This condition corresponds to the output of the average phone in a "normal" room environment.

Ensuring reliability does not seem to be a very difficult task. First of all, SNR can be improved by increasing the number of receiving microphones. On the other hand, in practical systems simple shielding is very easy to implement. And finally, even one error in recognition is not fatal: the user can always have another try. A recognizing device can easily identify cases of unsuccessful recognition and indicate the former case for the user to retry.

The recognition system can be implemented on any PC equipped with a sound card. The algorithm is so simple that the system can also be implemented as an embedded solution based on digital signal processors. Microphone requirements are not critical either: both the frequency response and SNR of entry level microphones are much better than those of mobile device ringers. This means that cheap stand-alone recognizers can be implemented and deployed at the points of service.

It is interesting to note that other devices capable of playing MIDI sequences (e.g., PDAs) can be used as well as mobile phones.

Measured transfer rate (40 bps) was considerably less than the estimation, obtained in our simple model—375 bps. We think that the reason for this was slow pulse decay rate in combination with non-linear frequency response. Amplitude of the note with frequency close to a local frequency response maximum might remain higher than amplitude of the consecutive note through the whole duration of the latter. Thus, the weaker sound of the second note might be not recognized.

However, we consider even such relatively slow transmission still suitable for the purposes of mobile authorization applications, because authorization data is usually small and its transmission time is not critical.

Our example (Figure 3) seems to be a quite practical situation—transmitting 11-symbol password during 2s is definitely not too long for a user. Typing the same token on the vending machine keyboard would easily take twice as long.

The acoustic presentation method might be an attractive feature for teenagers (e.g., mobile cinema tickets being one conceivable application).

ACKNOWLEDGMENTS

The authors would like to thank Petteri Koponen for the original idea.

REFERENCES

Khashchanskiy, V., & Kustov, A. (2001). Universal SIM Toolkit-based client for mobile authorization system. *Proceedings of the 3rd International Conference on Information Integration and Web-Based Applications & Services (IIWAS 2001)* (pp. 337-344).

Marple, S. Lawrence Jr. (1987). *Digital spectral analysis with applications.* Englewood Cliffs, NJ: Prentice-Hall.

3GPP TS 11.14. (1999). *Specification of the SIM application toolkit for the Subscriber Identity Module-Mobile Equipment (SIM-ME) interface.* Retrieved from http://www.3gpp.org/ftp/Specs/html-info/1114.htm

KEY TERMS

Fast Fourier Transform (FFT): An optimized form of the algorithm that calculates a complex spectrum of digitized signals. It is most widely used to obtain a so-called power spectrum as a square of a complex spectrum module. Power spectrum represents energy distribution along frequency axis.

Frequency Resolution: The minimum difference in frequencies which can be distinguished in a signal spectrum.

Frequency Response: For a device, circuit, or system, the ratio between output and input signal spectra.

Frequency Shift Keying (FSK): The digital modulation scheme that assigns fixed frequencies to certain bit sequences. Differential FSK (DFSK) uses frequency differences to encode bit sequences.

Harmonic Distortions: Alteration of the original signal shape caused by the appearance of higher harmonics of input signal at the output.

IR: Short-range infrared communication channel.

Musical Instrument Digital Interface (MIDI): A standard communications protocol that transfers musical notes between electronic musical instruments as sequences of events, like 'Note On', 'Note Off', and many others.

Sampling Frequency: The rate at which analogue signal is digitized by an analogue-to-digital converter (ADC) in order to convert the signal into numeric format that can be stored and processed by a computer.

Adaptive Transmission of Multimedia Data over UMTS

Antonios Alexiou
Patras University, Greece

Dimitrios Antonellis
Patras University, Greece

Christos Bouras
Patras University, Greece

INTRODUCTION

As communications technology is being developed, users' demand for multimedia services raises. Meanwhile, the Internet has enjoyed tremendous growth in recent years. Consequently, there is a great interest in using the IP-based networks to provide multimedia services. One of the most important areas in which the issues are being debated is the development of standards for the universal mobile telecommunications system (UMTS). UMTS constitutes the third generation of cellular wireless networks which aims to provide high-speed data access along with real-time voice calls. Wireless data is one of the major boosters of wireless communications and one of the main motivations of the next-generation standards.

Bandwidth is a valuable and limited resource for UMTS and every wireless network in general. Therefore, it is of extreme importance to exploit this resource in the most efficient way. Consequently, when a user experiences a streaming video, there should be enough bandwidth available at any time for any other application that the mobile user might need. In addition, when two different applications run together, the network should guarantee that there is no possibility for any of the above-mentioned applications to prevail against the other by taking all the available channel bandwidth. Since Internet applications adopt mainly TCP as the transport protocol, while streaming applications mainly use RTP, the network should guarantee that RTP does not prevail against the TCP traffic. This means that there should be enough bandwidth available in the wireless channel for the Internet applications to run properly.

BACKGROUND

Chen and Zachor (2004) propose a widely accepted rate control method in wired networks which is the equation-based rate control, also known as TFRC (TCP-friendly rate control). In this approach the authors use multiple TFRC connections as an end-to-end rate control solution for wireless streaming video. Another approach is presented by Fu and Liew (2003). As they mention, TCP Reno treats the occurrence of packet losses as a manifestation of network congestion. This assumption may not apply to networks with wireless channels, in which packet losses are often induced by noise, link error, or reasons other than network congestion. Equivalently, TCP Vegas uses queuing delay as a measure of congestion (Choe & Low, 2003). Thus, Fu and Liew (2003) propose an enhancement of TCP Reno and TCP Vegas for the wireless networks, namely TCP Veno.

Chen, Low, and Doyle (2005) present two algorithms that formulate resource allocation in wireless networks. These procedures constitute a preliminary step towards a systematic approach to jointly design TCP congestion control algorithms, not only to improve performance, but more importantly, to make interaction more transparent. Additionally, Xu, Tian, and Ansari (2005) study the performance characteristics of TCP New Reno, TCP SACK, TCP Veno, and TCP Westwood under the wireless network conditions and they propose a new TCP scheme, called TCP New Jersey, which is capable of distinguishing wireless packet losses from congestion.

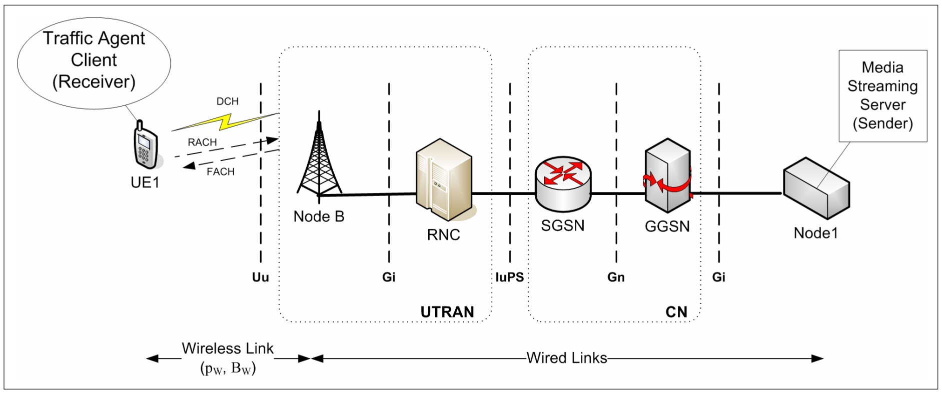
Recent work provides an overview of MPEG-4 video transmission over wireless networks (Zhao, Kok, & Ahmad, 2004). A critical issue is how we can ensure the QoS of video-based applications to be maintained at an acceptable level. Another point to consider is the unreliability of the network, especially of the wireless channels, because we observe packet losses resulting in a reduction of the video quality. The results demonstrate that the video quality can be substantially improved by preserving the high-priority video data during the transmission.

THE TCP-FRIENDLY RATE CONTROL PROTOCOL

TFRC is not actually a fully specified end to-end transmission protocol, but a congestion control mechanism that is designed

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Typical scenario for streaming video over UMTS


Traffic Agent Client (Receiver)
DCH
RACH
FACH
UE1
Node B
RNC
SGSN
GGSN
Node1
Media Streaming Server (Sender)
Uu
Gi
IuPS
Gn
Gi
UTRAN
CN
Wireless Link (p_W, B_W)
Wired Links

to operate fairly along with TCP traffic. Generally TFRC should be deployed with some existing transport protocol such as UDP or RTP in order to present its useful properties (Floyd, Handley, Padhye, & Widmer, 2000). The main idea behind TFRC is to provide a smooth transmission rate for streaming applications. The other properties of TFRC include slow response to congestion and the opportunity of not aggressively trying to make up with all available bandwidth. Consequently, in case of a single packet loss, TFRC does not halve its transmission rate like TCP, while on the other hand it does not respond rapidly to the changes in available network bandwidth. TFRC has also been designed to behave fairly when competing for the available bandwidth with concurrent TCP flows that comprise the majority of flows in today's networks. A widely popular model for TFRC is described by the following equation (Floyd & Fall, 1999):

$$T = \frac{kS}{RTT\sqrt{p}} \tag{1}$$

T represents the sending rate, *S* is the packet size, *RTT* is the end-to-end round trip time, *p* is the end-to-end packet loss rate, and *k* is a constant factor between 0.7 and 1.3 (Mahdavi & Floyd, 1997) depending on the particular derivation of equation (1).

The equation describes TFRC's sending rate as a function of the measured packet loss rate, round-trip time, and used packet size. More specifically, a potential congestion in the nodes of the path will cause an increment in the packet loss rate and in the round trip time according to the current packet size. Given this fluctuation, it is easy to determine the new transmission rate so as to avoid congestion and packet losses. Generally, TFRC's congestion control consists of the following mechanisms:

1. The receiver measures the packet loss event rate and feeds this information back to the sender.
2. The sender uses these feedback messages to calculate the round-trip-time (*RTT*) of the packets.
3. The loss event rate and the *RTT* are then fed into the TRFC rate calculation equation (described later in more detail) in order to find out the correct data sending rate.

ANALYSIS OF THE TFRC MECHANISM FOR UMTS

The typical scenario for streaming video over UMTS is shown in Figure 1, where the server is denoted by Node1 and the receiver by UE1. The addressed scenario comprises a UMTS radio cell covered by a Node B connected to an RNC. The model consists of a UE connected to DCH, as shown in Figure 1. In this case, the DCH is used for the transmission of the data over the air. DCH is a bi-directional channel and is reserved only for a single user. The common channels are the forward access channel (FACH) in the downlink and the random access channel (RACH) in the uplink.

The wireless link is assumed to have available bandwidth B_W, and packet loss rate p_w, caused by wireless channel error. This implies that the maximum throughput that could be achieved in the wireless link is $B_W (1 - p_W)$. There could also be packet loss caused by congestion at wired nodes denoted by $p_{node\ name}$ (node name: GGSN, SGSN, RNC, Node B). The end-to-end packet loss rate observed by the receiver is denoted as *p*. The streaming rate is denoted by *T*. This means that the streaming throughput is $T (1 - p)$. Under the above assumptions we characterize the wireless channel as underutilized if $T (1 - p) < B_W (1 - p_W)$. Given the above described scenario, the following are assumed:

1. The wireless link is the long-term bottleneck. This means that there is no congestion due to streaming traffic to the nodes GGSN, SGSN, and RNC.
2. There is no congestion at Node B due to the streaming application, if and only if the wireless bandwidth is underutilized—that is, $T (1 - p) < B_W (1 - p_W)$. This also implies that no queuing delay is caused at Node

B and hence, the round trip time for a given route has the minimum value (i.e., RTT_{min}). Thus, this assumption can be restated as follows: for a given route, $RTT = RTT_{min}$ if and only if $T(1 - p) \leq B_W(1 - p_W)$. This in turn implies that if $T(1 - p) > B_W(1 - p_W)$ then $RTT \geq RTT_{min}$.

3. The packet loss rate caused by wireless channel error (p_W) is random and varies from 0 to 0.16.
4. The backward route is error-free and congestion-free.

The communication between the sender and the receiver is based on RTP/RTCP sessions; and the sender, denoted by Node 1 (Figure 1), uses the RTP protocol to transmit the video stream. The client, denoted by UE1 (Figure 1), uses the RTCP protocol in order to exchange control messages. The mobile user in recurrent time space sends RTCP reports to the media server. These reports contain information about the current conditions of the wireless link during the transmission of the multimedia data between the server and the mobile user. The feedback information contains the following parameters:

- **Packet Loss Rate:** The receiver calculates the packet loss rate during the reception of sender data, based on RTP packets sequence numbers.
- **Timestamp of Every Packet Arrived at the Mobile User:** This parameter is used by the server for the RTT calculation of every packet.

The media server extracts the feedback information from the RTCP report and passes it through an appropriate filter. The use of filter is essential for the operation of the mechanism in order to avoid wrong estimations of the network conditions. On the sender side, the media server using the feedback information estimates the appropriate rate of the streaming video so as to avoid network congestion. The appropriate transmission rate of the video sequence is calculated from equation (1), and the media server is responsible for adjusting the sending rate with the calculated value. Obviously, the media server does not have the opportunity to transmit the video in all the calculated sending rates. However, it provides a small variety of them and has to approximate the calculated value choosing the sending rate from the provided transmission rates.

This extends the functionality of the whole congestion control mechanism. More specifically, the sender does not have to change the transmission rate every time it calculates a new one with a slight difference from the previous value. Consequently, it changes the transmission rate of the multimedia data to one of the available sending rates of the media server as has already been mentioned. In this approach, the number of the changes in the sending rate is small and the mobile user does not deal with a continually different transmission rate.

As mentioned above, it is essential to keep a history of the previous calculated values for the transmission rate. Having this information, the media server can estimate the smoothed transmission rate, using the *m* most recent values of the calculated sending rate from the following equation:

$$T^{Smoothed} = \frac{\sum_{i=1}^{m} w_i \cdot T_{m+1-i}^{Smoothed}}{\sum_{i=1}^{m} w_i} \qquad (2)$$

The value *m*, used in calculating the transmission rate, determines TFRC's speed in responding to changes in the level of congestion (Handley, Floyd, Padhye, & Widmer, 2003). The weights w_i are appropriately chosen so that the most recent calculated sending rates receive the same high weights, while the weights gradually decrease to 0 for older calculated values.

Equivalently to the calculation of the transmission rate, the mobile user (client) measures the packet loss rate p_1 based on the RTP packets sequence numbers. This information is sent to the media server via the RTCP reports. In order to prevent a single spurious packet loss having an excessive effect on the packet loss estimation, the server smoothes the values of packet loss rate using the filter of the following equation, which computes the weighted average of the *m* most recent loss rate values (Vicisiano, Rizzo, & Crowcroft, 1998).

$$p_1^{Smoothed} = \frac{\sum_{i=1}^{m} w_i \cdot p_{1,m+1-i}^{Smoothed}}{\sum_{i=1}^{m} w_i} \qquad (3)$$

The value of $p_1^{Smoothed}$ is then used by equation (1) for the estimation of the transmission rate of the multimedia data. The weights w_i are chosen as in the transmission rate estimation.

FUTURE TRENDS

The most prominent enhancement of the adaptive real-time applications is the use of multicast transmission of the multimedia data. The multicast transmission of multimedia data has to accommodate clients with heterogeneous data reception capabilities. To accommodate heterogeneity, the server may transmit one multicast stream and determine the transmission rate that satisfies most of the clients (Byers et al., 2000). Additionally, Vickers, Albuquerque, and Suda (1998) present different approaches where the server transmits multiple multicast streams with different transmission rates allocating the client at these streams, as well as using layered encoding and transmitting each layer to a different multicast stream. An interesting survey of techniques for multicast multimedia data over the Internet is presented in Li, Ammar, and Paul (1999).

Single multicast stream approaches have the disadvantage that clients with a low-bandwidth link will always get a high-bandwidth stream if most of the other members are connected via a high-bandwidth link, and the same is true the other way around. This problem can be overcome with the use of a multi-stream multicast approach. Single multicast stream approaches have the advantages of easy encoder and decoder implementation and simple protocol operation, due to the fact that during the single multicast stream approach, there is no need for synchronization of clients' actions (as the multiple multicast streams and layered encoding approaches require).

The subject of adaptive multicast of multimedia data over networks with the use of one multicast stream has engaged researchers all over the world. During the adaptive multicast transmission of multimedia data in a single multicast stream, the server must select the transmission rate that satisfies most the clients with the current network conditions. Totally, three approaches can be found in the literature for the implementation of the adaptation protocol in a single stream multicast mechanism: equation based (Rizzo, 2000; Widmer & Handley, 2001), network feedback based (Byers et al., 2000), or a combination of the above two approaches (Sisalem & Wolisz, 2000).

CONCLUSION

An analysis of the TCP friendly rate control mechanism for UMTS has been presented. The TFRC mechanism gives the opportunity to estimate the appropriate transmission rate of the video data for avoiding congestion in the network. The three goals of this rate control could be stated as follows. First, the streaming rate does not cause any network instability (i.e., congestion collapse). Second, TFRC is assumed to be TCP Friendly, which means that any application that transmits data over a network presents friendly behavior towards the other flows that coexist in the network and especially towards the TCP flows that comprise the majority of flows in today's networks. Third, it leads to the optimal performance—that is, it results in the highest possible throughput and lowest possible packet loss rate. Furthermore, an overview of video transmission over UMTS using real-time protocols such as RTP/RTCP has been presented.

REFERENCES

Byers, J., Frumin, M., Horn, G., Luby, M., Mitzenmacher, M., Roetter, A., & Shaver, W. (2000). FLID-DL: Congestion control for layered multicast. *Proceedings of the International Workshop on Networked Group Communication* (pp. 71-81), Palo Alto, CA.

Chen, M., & Zachor, A. (2004). Rate control for streaming video over wireless. *IEEE INFOCOM,* Hong Kong, China, (pp. 1181-1190).

Chen, L., Low, S., & Doyle, J. (2005). Joint congestion control and media access control design for ad hoc wireless networks. *IEEE INFOCOM*, Miami, FL.

Choe, H., & Low, S. (2003). Stabilized Vegas. *IEEE INFOCOM, 22*(1), 2290-2300.

Floyd, S., Handley, M., Padhye, J., & Widmer, J. (2000). Equation-based congestion control for unicast applications. *Proceedings of ACM SIGCOMM,* Stockholm, Sweden, (pp. 43-56).

Floyd, S., & Fall, K. (1999). Promoting the use of end-to-end congestion control in the Internet. *IEEE/ACM Transactions on Networking, 7*(4), 458-472.

Fu, C. P., & Liew, S. C. (2003). TCP Veno: TCP enhancement for transmission over wireless access networks. *IEEE Journal on Selected Areas in Communications, 21*(2), 216-228.

Handley, M., Floyd, S., Padhye, J., & Widmer, J. (2003). TCP Friendly Rate Control (TFRC). *RFC, 3448.*

Li, X., Ammar, M., & Paul, S. (1999). Video multicast over the Internet. *IEEE Network Magazine, 12*(2), 46-60.

Mahdavi, J., & Floyd, S. (1997). *TCP-Friendly unicast rate-based flow control.* Retrieved from http://www.psc.edu/networking/papers/tcp_friendly.html

Rizzo, L. (2000). pgmcc: A TCP-friendly single-rate multicast congestion control scheme. *Proceedings of ACM SIGCOMM,* Stockholm, Sweden, (pp. 17-28).

Sisalem, D., & Wolisz, A. (2000). LDA+ TCP-Friendly adaptation: A measurement and comparison study. *Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, Chapel Hill, NC.

Vicisiano, L., Rizzo, L., & Crowcroft, J. (1998). TCP-like congestion control for layered multicast data transfer. *IEEE INFOCOM*, San Francisco, CA, (pp. 996-1003).

Vickers, J., Albuquerque, N., & Suda, T. (1998). Adaptive multicast of multi-layered video: Rate-based and credit-based approaches. *IEEE INFOCOM*, San Francisco, CA, (pp. 1073-1083).

Widmer, J., & Handley, M. (2001). Extending equation-based congestion control to multicast applications. *Proceedings of ACM SIGCOMM*, San Diego, CA, (pp. 275-286).

Xu, K., Tian, Y., & Ansari, N. (2005). Improving TCP performance in integrated wireless communications networks. *Computer Networks, Science Direct, 47*(2), 219-237.

Zhao, J., Kok, C., & Ahmad, I. (2004). MPEG-4 video transmission over wireless networks: A link level performance study. *Wireless Networks, 10*(2), 133-146.

KEY TERMS

Adaptive Real-Time Application: An application that has the capability to transmit multimedia data over heterogeneous networks and adapt media transmission to network changes.

Delay Jitter: The mean deviation (smoothed absolute value) of the difference in packet spacing at the receiver compared to the sender for a pair of packets.

Frame Rate: The rate of the frames, which are encoded by video encoder.

Multimedia Data: Data that consist of various media types like text, audio, video, and animation.

Packet Loss Rate: The fraction of the total transmitted packets that did not arrive at the receiver.

RTP/RTCP: Protocol used for the transmission of multimedia data. The RTP performs the actual transmission, and the RTCP is the control and monitoring transmission.

Addressing the Credibility of Mobile Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

Mobile access has opened new vistas for various sectors of society including businesses. The ability that anyone using (virtually) any device could be reached anytime and anywhere presents a tremendous commercial potential. Indeed, the number of mobile applications has seen a tremendous growth in the last few years.

In retrospect, the fact that almost *anyone* can set up a mobile application claiming to offer products and services raises the question of credibility from a consumer's viewpoint. The obligation of establishing credibility is essential for an organization's reputation (Gibson, 2002) and for building consumers' trust (Kamthan, 1999). If not addressed, there is a potential for lost consumer confidence, thus significantly reducing the advantages and opportunities the mobile Web as a medium offers. If a mobile application is not seen as credible, we face the inevitable consequence of a product, however functionally superior it might be, rendered socially isolated.

The rest of the article is organized as follows. We first provide the motivational background necessary for later discussion. This is followed by introduction of a framework within which different types of credibility in the context of mobile applications can be systematically addressed and thereby improved. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

In this section, we present the fundamental concepts underlying credibility, and present the motivation and related work for addressing credibility within the context of mobile applications.

Basic Credibility Concepts

For the purposes of this article, we will consider credibility to be synonymous to (and therefore interchangeable with) believability (Hovland, Janis, & Kelley, 1953). We follow the terminology of Fogg and Tseng (1999), and view credibility and trust as being slightly different. Since trust indicates a *positive* belief about a person, object, or process, we do not consider credibility and trust to be synonymous.

It has been pointed out in various studies (Fogg, 2003; Metzger, 2005) that credibility consists of two primary dimensions, namely *trustworthiness* and *expertise* of the source of some information. Trustworthiness is defined by the terms such as well-intentioned, truthful, unbiased, and so on. The trustworthiness dimension of credibility captures the *perceived* goodness or morality of the source. Expertise is defined by terms such as knowledgeable, experienced, competent, and so on. The expertise dimension of credibility captures the *perceived* knowledge and skill of the source. Together, they suggest that "highly credible" mobile applications will be perceived to have high levels of *both* trustworthiness and expertise.

We note that trustworthiness and expertise are at such a high level of abstraction that direct treatment of any of them is difficult. Therefore, in order to improve credibility, we need to find quantifiable attributes that can improve each of these dimensions.

A Classification of Credibility

The following taxonomy helps associating the concept of credibility with a specific user class in context of a mobile application. A user could consider a mobile application to be credible based upon direct interaction with the application (*active credibility*), or consider it to be credible in absence of any direct interaction but based on certain pre-determined notions (*passive credibility*). Based on the classification of credibility in computer use (Fogg & Tseng, 1999) and adapting them to the domain of mobile applications, we can decompose these further.

There can be two types of *active credibility:* (1) *surface credibility,* which describes how much the user believes the mobile application is based on simple inspection; and (2) *experienced credibility,* which describes how much the user believes the mobile application is based on first-hand experience in the past.

There can be two types of *passive credibility:* (1) *presumed credibility,* which describes how much the user believes the mobile application because of general assumptions that the user holds; and (2) *reputed credibility,* which describes how

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

much the user believes the mobile application because of a reference from a third party.

Finally, credibility is not absolute with respect to users and with respect to the application itself (Metzger, Flanagin, Eyal, Lemus, & McCann, 2003). Also, credibility can be associated with a whole mobile application or a part of a mobile application. For example, a user may question the credibility of information on a specific product displayed in a mobile application. We contend that for a mobile application to be labeled non-credible, there must exist at least a part of it that is labeled non-credible based on the above classification by at least one user.

The Origins and Significance of the Problem of Mobile Credibility

The credibility of mobile applications deserves special attention for the following reasons:

- **Delivery Context:** Mobile applications are different from the desktop or Web environments (Paavilainen, 2002) where context-awareness (Sadeh, Chan, Van, Kwon, & Takizawa, 2003) is a unique challenge. The delivery context in a changing environment of mobile markup languages, variations in user agents, and constrained capabilities of mobile devices presents unique challenges towards active credibility.
- **Legal Context:** Since the stakeholders of a mobile application need not be co-located (different jurisdictions in the same country or in different countries), the laws that govern the provider and the user may be different. Also, the possibilities of fraud such as computer domain name impersonation (commonly known as "pharming") or user identity theft (commonly known as "phishing") with little legal repercussions for the perpetrators is relatively high in a networked environment. These possibilities can impact negatively on presumed credibility.
- **User Context:** Users may deploy mobile devices with varying configurations, and in the event of problems with a mobile service, may first question the provider rather than the device that they own. In order for providers of mobile portals to deliver user-specific information and services, they need to know details about the user (such as profile information, location, and so on). This creates the classical dichotomy between personalization and privacy, and striking a balance between the two is a constant struggle for businesses (Kasanoff, 2002). The benefits of respecting one can adversely affect the other, thereby impacting their credibility in the view of their customers. Furthermore, the absence of a human component from non-proximity or "facelessness" of the provider can shake customer confidence and create negative perceptions in a time of crisis such as denial of service or user agent crash. These instances can lead to a negative passive credibility.

Initiatives for Improving Mobile Credibility

There have been initiatives to address the credibility of Web applications such as a user survey to identify the characteristics that users consider necessary for a Web application to be credible (Fogg et al., 2001) and a set of guidelines (Fogg, 2003) for addressing *surface, experienced, presumed,* and *reputed credibility* of Web applications.

However, these efforts are limited by one or more of the following issues. The approach towards ensuring and/or evaluating credibility is not systematic, the proposed means for ensuring credibility is singular (only guidelines), and the issue of feasibility of the means is not addressed. Moreover, these guidelines are not specific to mobility, are not prioritized and the possibility that they can contradict each other is not considered, can be open to broad interpretation, and are stated at such a high level that they may be difficult to realize by a novice user.

ADDRESSING THE CREDIBILITY OF MOBILE APPLICATIONS

In this section, we consider approaches for understanding and improving active credibility of mobile applications.

A Framework for Addressing Active Credibility of Mobile Applications

To systematically address the active credibility of mobile applications, we take the following steps:

1. View credibility as a qualitative aspect and address it indirectly via quantitative means.
2. Select a theoretical basis for communication of information (semiotics), and place credibility in its setting.
3. Address semiotic quality in a practical manner.

Based on this and using the primary dimensions that affect credibility, we propose a framework for active credibility of mobile applications (see Table 1). The external attributes (denoted by E) are extrinsic to the software product and are directly a user's concern, while internal attributes (denoted by I) are intrinsic to the software product and are directly an engineer's concern. Since not all attributes corresponding to a semiotic level are at the same echelon, the different tiers are denoted by "Tn."

Table 1. A semiotic framework for active credibility of mobile applications

Semiotic Level	**Quality Attributes**	**Means for Credibility Assurance and Evaluation**	**Decision Support**
Social	**Credibility**	- “Expert” Knowledge (Principles, Guidelines, Patterns) - Inspections - Testing - Metrics - Tools	Feasibility
	Aesthetics, Legality, Privacy, Security, (Provider) Transparency [T5;E]		
Pragmatic	Accessibility, Usability [T4;E]		
	Interoperability, Portability, Reliability, Robustness [T3;E]		
Semantic	Completeness and Validity [T2;I]		
Syntactic	Correctness [T1;I]		

We now describe each of the components of the framework in detail.

Semiotic Levels

The first column of Table 1 addresses semiotic levels. Semiotics (Stamper, 1992) is concerned with the use of symbols to convey knowledge.

From a semiotics perspective, a representation can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social, each depending on the previous one in that order. The physical and empirical levels are concerned with the physical representation of signs in hardware and communication properties of signs, and are not of direct concern here. The syntactic level is responsible for the formal or structural relations between signs. The semantic level is responsible for the relationship of signs to what they stand for. The pragmatic level is responsible for the relation of signs to interpreters. The social level is responsible for the manifestation of social interaction with respect to signs.

Quality Attributes

The second column of Table 1 draws the relationship between semiotic levels and corresponding quality attributes.

Credibility belongs to the social level and depends on the layers beneath it. The external quality attributes *legality, privacy, security,* and *(provider) transparency* also at the social level depend upon the external quality attributes *accessibility* and *usability* at the pragmatic level, which in turn depend upon the external quality attributes *interoperability, performance, portability, reliability,* and *robustness* also at the pragmatic level. (We note here that although *accessibility* and *usability* do overlap in their design and implementation, they are not identical in their goals for their user groups.)

We discuss in some detail only the entries in the social level. Aesthetics is close to human senses and perception, and plays a crucial role in making a mobile application “salient” to its customers beyond simply the functionality it offers. It is critical that the mobile application be legal (e.g., is legal in the jurisdiction it operates and all components it makes use of are legal); takes steps to respect a user’s privacy (e.g., does not use or share user-supplied information outside the permitted realm); and be secure (e.g., in situations where financial transactions are made). The provider must take all steps to be transparent with respect to the user (e.g., not include misleading information such as the features of products or services offered, clearly label promotional content, make policies regarding returning/exchanging products open, and so on).

The internal quality attributes for syntactic and semantic levels are inspired by Lindland, Sindre, and Sølvberg (1994) and Fenton and Pfleeger (1997). At the semantic level, we are only concerned with the conformance of the mobile application to the domain(s) it represents (that is, semantic correctness or completeness) and vice versa (that is, semantic validity). At the syntactic level the interest is in conformance with respect to the languages used to produce the mobile application (that is, syntactic correctness).

The definitions of each of these attributes can vary in the literature, and therefore it is important that they be adopted and followed consistently. For example, the definition of usability varies significantly across ISO/IEC Standard 9126 and ISO Standard 9241 with respect to the perspective taken in their formulation.

Means for Credibility Assurance and Evaluation

The third column of Table 1 lists (in no particular order, by no means complete, and not necessarily mutually exclusive) the means for assuring and evaluating active credibility.

- **"Expert" Body of Knowledge:** The three types of knowledge that we are interested in are principles, guidelines, and patterns. Following the basic principles (Ghezzi, Jazayeri, & Mandrioli, 2003) underlying a mobile application enables a provider to improve quality attributes related to (T1-T3) of the framework. The guidelines encourage the use of conventions and good practice, and could also serve as a checklist with respect to which an application could be heuristically or otherwise evaluated. There are guidelines available for addressing accessibility (Chisholm, Vanderheiden, & Jacobs, 1999; Ahonen, 2003), security (McGraw & Felten, 1998), and usability (Bertini, Catarci, Kimani, & Dix, 2005) of mobile applications. However, guidelines tend to be more useful for those with an expert knowledge than for a novice to whom they may seem rather general to be of much practical use. Patterns are reusable entities of knowledge and experience aggregated by experts over years of "best practices" in solving recurring problems in a domain including that in mobile applications (Roth, 2001, 2002). They are relatively more structured compared to guidelines and provide better opportunities for sharing and reuse. There is, however, a lack of patterns that clearly address quality concerns in mobile applications. Also, there is a cost of adaptation of patterns to new contexts.
- **Inspections:** Inspections (Wiegers, 2002) are a rigorous form of auditing based upon peer review that can address quality concerns at both technical and social levels (T1-T5), and help improve the credibility of mobile applications. Inspections could, for example, decide what information is and is not considered "promotional," help improve the labels used to provide cues to a user (say, in a navigation system), and assess the readability of documents. Still, inspections do involve an initial cost overhead from training each participant in the structured review process, and the logistics of checklists, forms, and reports.
- **Testing:** Some form of testing is usually an integral part of most development models of mobile applications (Nguyen, Johnson, & Hackett, 2003). There are test suites and test harnesses for many of the languages commonly used for representation of information in mobile applications. However, due to its very nature, testing addresses quality concerns of only some of the technical and social levels (T1, subset of T2, T3, T4, subset of T5). Therefore, testing *complements* but does not replace inspections. Accessibility or usability testing that requires hiring real users, infrastructure with video monitoring, and subsequent analysis of data can prove to be prohibitive for small-to-medium-size enterprises.
- **Metrics:** In a resource-constrained environment of mobile devices, efficient use of time and space is critical. Metrics (Fenton & Pfleeger, 1997) provide a quantitative means for making qualitative judgments about quality concerns at technical levels. For example, metrics for a document or image size can help compare and make a choice between two designs, or metrics for structural complexity could help determine the number of steps required in navigation, which in turn could be used to estimate user effort. However, well-tested metrics for mobile applications are currently lacking. We also note that a dedicated use of metrics on a large scale usually requires tool support.
- **Tools:** Tools that have help improve quality concerns at technical and social levels. For example, tools can help engineers detect security breaches, report violations of accessibility or usability guidelines, find non-conformance to markup or scripting language syntax, suggest image sizes favorable to the small devices, or detect broken links. However, at times, tools cannot address some of the technical quality concerns (like complete semantic correctness of the application with respect to the application domain), as well as certain social quality concerns (like provider intent or user bias). Therefore, the use of tools as means for automatic quality assurance or evaluation should be kept in perspective.

Decision Support

A mobile application project must take a variety of constraints into account: organizational constraints of time and resources (personnel, infrastructure, budget, and so on) and external forces (market value, competitors, and so on). These compel providers to make quality-related decisions that, apart from being sensitive to credibility, must also be feasible.

For example, the provider of a mobile application should carry out intensive accessibility and usability evaluations, but ultimately that application must be delivered on a timely basis. Also, the impossibility of complete testing is well known.

Indeed, the last column of Table 1 acknowledges that with respect to any assurance and/or evaluation, and includes feasibility as an all-encompassing consideration on the layers to make the framework practical. There are well-known techniques such as analytical hierarchy process (AHP) and quality function deployment (QFD) for carrying out feasibility analysis, and further discussion of this aspect is beyond the scope of this article.

Limitations of Addressing Credibility

We note here that credibility, as is reflected by its primary dimensions, is a socio-cognitive concern that is not always amenable to a purely technological treatment. However, by decomposing it into quantifiable elements and approaching them in a systematic and feasible manner, we can make improvements towards its establishment.

We assert that the quality attributes we mention in pragmatic and social levels are necessary but make no claim of their sufficiency. Indeed, as we move from bottom to top, the framework gets less technically oriented and more human oriented. Therefore, finding sufficient conditions for establishing credibility is likely to be an open question, and it may be virtually impossible to provide complete guarantees for credibility.

FUTURE TRENDS

In the previous section, we discussed active credibility; the issue of passive credibility poses special challenges and is a potential area of future research. We now briefly look at the case of reputed credibility.

In case of Web applications, there have been two notable initiatives in the direction of addressing reputed credibility, namely WebTrust and TRUSTe. In response to the concerns related to for business-to-consumer electronic commerce and to increase consumer confidence, the American Institute of Certified Public Accountants (AICPA) and Canadian Institute of Chartered Accountants (CICA) have developed WebTrust Principles and Criteria and the related WebTrust seal of assurance. Independent and objective certified public accountant or chartered accountants, who are licensed by the AICPA or CICA, can provide assurance services to evaluate and test whether a particular Web application meets these principles and criteria. The TRUSTe program enables companies to develop privacy statements that reflect the information gathering and dissemination practices of their Web application. The program is equipped with the TRUSTe "trustmark" seal that takes users directly to a provider's privacy statement. The trustmark is awarded only to those that adhere to TRUSTe's established privacy principles and agree to comply with ongoing TRUSTe oversight and resolution process. Admittedly, not in the realm of pure academia, having similar quality assurance and evaluation programs for mobile applications, and perhaps even the use of ISO 9001:2000 as a basis for a certification, would be of interest.

A natural extension of the preceding discussion on credibility could be in the context of the next generation of mobile applications such as semantic mobile applications (Alesso & Smith, 2002) and mobile Web services (Salmre, 2005). For example, ontological representation of information can present certain human-centric challenges (Kamthan & Pai, 2006) that need to be overcome for it to be a credible knowledge base.

Finally, viewing a mobile application as an information system, it would of interest to draw connections between credibility and ethics (Johnson, 1997; Tavani, 2004).

CONCLUSION

Although there have been significant advances towards enabling the technological infrastructure (Coyle, 2001) for mobile access in the past decade, there is much to be done in addressing the social challenges. Addressing credibility of mobile applications in a systematic manner is one step in that direction.

The organizations that value credibility of their mobile applications need to take two aspects into consideration: (1) take a *systematic* approach to the development of the mobile applications, and (2) consider credibility as a first-class concern *throughout* the process. The former need to particularly include support for modeling a user's environment (context, task, and device) (Gandon & Sadeh, 2004) and mobile user interface engineering. The latter implies that credibility is viewed as a *mandatory* non-functional requirement during the analysis phase and treated as a central design concern in the synthesis phase.

In a user-centric approach to engineering, mobile applications belong to an *ecosystem* that includes both the people and the product. If the success of a mobile application is measured by use of its services, then establishing credibility with the users is critical for the providers. By making efforts towards improving the criteria that directly or indirectly affect credibility, the providers can meet user expectations and change the user perceptions in their favor.

REFERENCES

Ahonen, M. (2003, September 19). Accessibility challenges with mobile lifelong learning tools and related collaboration. *Proceedings of the Workshop on Ubiquitous and Mobile Computing for Educational Communities (UMOCEC 2003)*, Amsterdam, The Netherlands.

Alesso, H. P., & Smith, C. F. (2002). *The intelligent wireless Web.* Boston: Addison-Wesley.

Bertini, E., Catarci, T., Kimani, S., & Dix, A. (2005). A review of standard usability principles in the context of mobile computing. *Studies in Communication Sciences, 1*(5), 111-126.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0.* W3C Recommendation, World Wide Web Consortium (W3C).

Coyle, F. (2001). *Wireless Web: A manager's guide.* Boston: Addison-Wesley.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software metrics: A rigorous & practical approach.* International Thomson Computer Press.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do.* San Francisco: Morgan Kaufmann.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al. (2001, March 31-April 5). What makes Web sites credible?: A report on a large quantitative study. *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference,* Seattle, WA.

Fogg, B. J., & Tseng, S. (1999, May 15-20). The elements of computer credibility. *Proceedings of the ACM CHI 99 Conference on Human Factors in Computing Systems,* Pittsburgh, PA.

Gandon, F. L., & Sadeh, N. M. (2004, June 1-3). Context-awareness, privacy and mobile access: A Web semantic and multiagent approach. *Proceedings of the 1st French-Speaking Conference on Mobility and Ubiquity Computing,* Nice, France (pp. 123-130).

Gibson, D. A. (2002). *Communities and reputation on the Web.* PhD Thesis, University of California, USA.

Ghezzi, C., Jazayeri, M., & Mandrioli, D. (2003). *Fundamentals of software engineering* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hovland, C. I., Janis, I. L., & Kelley, J. J. (1953). *Communication and persuasion.* New Haven, CT: Yale University Press.

Johnson, D. G. (1997). Ethics online. *Communications of the ACM, 40*(1), 60-65.

Lindland, O. I., Sindre, G., & Sølvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE Software, 11*(2), 42-49.

Kamthan, P. (1999). *E-commerce on the WWW: A matter of trust.* Internet Related Technologies (IRT.ORG).

Kamthan, P., & Pai, H.-I. (2006, May 21-24). Human-centric challenges in ontology engineering for the semantic Web: A perspective from patterns ontology. *Proceedings of the 17th Annual Information Resources Management Association International Conference (IRMA 2006)*, Washington, DC.

Kasanoff, B. (2002). *Making it personal: How to profit from personalization without invading privacy.* New York: John Wiley & Sons.

McGraw, G., & Felten, E. W. (1998). Mobile code and security. *IEEE Internet Computing, 2*(6).

Metzger, M. J. (2005, April 11-13). Understanding how Internet users make sense of credibility: A review of the state of our knowledge and recommendations for theory, policy, and practice, *Proceedings of the Internet Credibility and the User Symposium,* Seattle, WA.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D., & McCann, R. (2003). Bringing the concept of credibility into the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication Yearbook, 27,* 293-335.

Nguyen, H. Q., Johnson, R., & Hackett, M. (2003). *Testing applications on the Web: Test planning for mobile and Internet-based systems* (2nd ed.). New York: John Wiley & Sons.

Paavilainen, J. (2002). *Mobile business strategies: Understanding the technologies and opportunities.* Boston: Addison-Wesley.

Roth, J. (2001, September 10). Patterns of mobile interaction. *Proceedings of the 3rd International Workshop on Human Computer Interaction with Mobile Devices (Mobile HCI 2001)*, Lille, France.

Roth, J. (2002). Patterns of mobile interaction. *Personal and Ubiquitous Computing, 6*(4), 282-289.

Sadeh, N. M., Chan, T.-C., Van, L., Kwon, O., & Takizawa, K. (2003, June 9-12). A semantic Web environment for context-aware m-commerce. *Proceedings of the 4th ACM Conference on Electronic Commerce,* San Diego, CA (pp. 268-269).

Salmre, I. (2005). *Writing mobile code: Essential software engineering for building mobile applications.* Boston: Addison-Wesley.

Stamper, R. (1992, October 5-8). Signs, organizations, norms and information systems. *Proceedings of the 3rd Australian Conference on Information Systems*, Wollongong, Australia.

Tavani, H. T. (2004). *Ethics and technology: Ethical issues in an age of information and communication technology.* New York: John Wiley & Sons.

Wiegers, K. (2002). *Peer reviews in software: A practical guide.* Boston: Addison-Wesley.

KEY TERMS

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Mobile Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering, and management principles, and disciplined and systematic approaches to the successful development, deployment, and maintenance of high-quality mobile Web applications.

Personalization: A strategy that enables delivery that is customized to the user and user's environment.

Quality: The totality of features and characteristics of a product or a service that bear on its ability to satisfy stated or implied needs.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and their representations.

User Profile: A information container describing user needs, goals, and preferences.

Adoption and Diffusion of M-Commerce

Ranjan B. Kini
Indiana University Northwest, USA

Subir K. Bandyopadhyay
Indiana University Northwest, USA

INTRODUCTION

Mobile commerce (or in short, m-commerce) is currently at the stage where e-commerce was a decade ago. Many of the concerns consumers had regarding e-commerce (such as security, confidentiality, and reliability) are now directed towards m-commerce. To complicate the matter further, the lack of a standardized technology has made m-commerce grow in multiple directions in different parts of the world. Thus, the popularity of m-commerce-based services varies by country, by culture, and by individual user. For example, in Europe the most popular application is SMS (short message service) or text messaging, in Japan interactive games and picture exchange via NTT DoCoMo i-mode, and in North America e-mail via interactive pagers (such as RIM BlackBerry) and wireless application protocol-based (WAP-based) wireless data portals providing news, stock quotes, and weather information. It is safe to predict that these applications will take on different forms as the technologies mature, devices become more capable in form and functionality, and service providers become more innovative in their business models.

It is true that m-commerce has witnessed spectacular growth across the globe. It is also encouraging that several factors are expected to accelerate the pace of adoption of m-commerce. Notable among these drivers is convergence in the voice/data industry, leaping improvements in related technology and standards, adoptive technology culture in many parts of the world, and governmental and regulatory initiatives.

Despite the undisputed promise of m-commerce, there are several barriers that are slowing the pace of adoption of m-commerce. The major barriers include: (a) lack of good business models to generate revenues, (b) perception of lack of security, (c) short product lifecycle due to rapidly changing technology, (d) non-convergence of standards, (e) usability of devices, (f) limitation of bandwidth, and (g) cost.

Many of the aforesaid were common to e-commerce also at its introduction and growth stage. We strongly believe it is worthwhile to investigate how e-commerce has been able to overcome these barriers so that we can incorporate some of the successful strategies to m-commerce. In our study, we will first compare and contrast e-commerce and m-commerce with respect to a set of common criteria such as: (1) hardware requirement, (2) software requirement, (3) connection or access, and (4) content. In the process, we will identify the principal barriers to the development of m-commerce as outlined in the above list.

The Growth in E-Commerce

Electronic commerce or e-commerce is the mode of commerce wherein the communication and transactions related to marketing, distributing, billing, communicating, and payment related to exchange of goods or services is conducted through the Internet, communication networks, and computers. Since the Department of Defense opened up the Internet for the public to access in 1991, there has been exponential growth in the number of Web sites, users on the Web, commerce through the Web, and now change of lifestyle through the Web (Pew, 2006).

The chronology of events shows that as the Internet became easier and cheaper to use, and as the applications (such as e-mail and Web interaction) became necessary or useful to have, the rate of adoption of the Internet accelerated. In fact, the rate of adoption of the Internet surpassed all projections that were made based on the traditional technology adoption rates that were documented for electricity, automobile, radio, telephone, and television (Pew, 2006). Unfortunately, the over-enthusiastic media hyped up the growth rate to an unsustainable level, leading to unprecedented growth of investment in the Internet technologies and followed by a melt-down in the stock market. This shattered the confidence in Internet technologies in the investment market. Although there was a significant deceleration in IT investment, e-commerce has rebounded to a large extent since the dot.com bust. It has been growing at about 30% compound rate per year (Pew, 2006).

In the last 10 years, the adoption of e-commerce has been extensively studied both by academicians as well as practitioners. During this period e-commerce and the scope of its definition also went through various iterations. For example, people may not buy a car on the Internet, but it is documented that 65% of car buyers have done extensive research on the Web about the car they eventually buy. Is this e-commerce? Should we restrict the e-commerce definition

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

to financial exchange for goods or services? We have various such examples in the marketplace where extensive research about the product or service is conducted on the Internet, but the final purchase is made in the physical environment. Hence, although the number of consumer financial transactions has not grown to the level industry projected initially, there has been a significantly high rate of adoption of the activities supporting e-commerce.

In addition, there has been a very high rate of adoption of business-to-business (B2B) commerce both in terms of financial and supporting transactions. In this article, we are interested in business-to-consumer (B2C) commerce. Hence, the comparison and contrast is made between e-commerce and m-commerce. All our discussion henceforth will be on B2C commerce using desktop and/or mobile technologies.

The Growth Potential of M-Commerce

Mobile commerce is the model of commerce that performs transactions using a wireless device and data connection that result in the transfer of value in exchange for information, services, or goods. Mobile commerce is facilitated generally by mobile phones and newly developed handheld devices. It includes services such as banking, payment, ticketing, and other related services (DEVX, 2006; Kini & Thanarithiporn, 2005).

Currently, most m-commerce activity is performed using mobile phones or handsets. This type of commerce is common in Asian countries led by Japan and South Korea. Industry observers are expecting that the United States will catch up soon, with mobile phones replacing existing devices such as ExxonMobil's Speedpass (eMarketer, 2005; Kini & Thanarithiporn, 2005).

Although the U.S. is lagging behind many countries in Asia and Europe in m-commerce, a UK-based research firm projects North American m-commerce users to total 12 million by 2009, with two-thirds of them using the devices to buy external items such as tickets and goods, and a third of them using it to make smaller transactions through vending machines (eMarketer, 2005). The firm also notes that there is a large potential number of the 95 million current American teens who are already making purchases on the Web that will adopt m-commerce. However, the study also remarks that generating widespread user interest in m-commerce and addressing security fears of mobile payment technologies and m-commerce services are critical in achieving a high level of adoption (eMarketer, 2005).

While the Asia Pacific Research Group (APRG, 2006) projected in 2002 that global m-commerce would reach US$10 billion 2005, Juniper Research currently projects that the global mobile commerce market, comprising mobile entertainment downloads, ticket purchases, and point-of-sale (POS) transactions, will grow to $88 billion by 2009, largely on the strength of micro-payments (e.g., vending machine type purchases). See eMarketer (2005) for more details.

Today, a large percentage of mobile phone users use mobile phones to download ring tones and play games; hence content-based m-commerce is expected to make up a small percentage of m-commerce. One recent study, however, projects that in the future mobile phone users will move up the value chain from purchases that are used and enjoyed on the mobile phone to external items such as tickets, snacks, public transportation, newspapers, and magazines (eMarketer, 2005).

Diffusion Models of Technology Adoption

There are many models that have been formulated and studied with regard to technology adoption, acceptance, diffusion, and continued adoption. These theories identify factors that are necessary to support different levels of adoption of information and communication technologies (ICTs). Notable among these models are the innovation-diffusion theory (Roger, 1995), technology acceptance model (or TAM) based on the theory of reasoned action (Davis, 1989; Fishbein & Ajzen, 1975), extended TAM2 model that incorporates social factors (Venkatesh & Davis, 2000), technology adoption model based on the theory of planned behavior (Ajzen & Fishbein, 1980), post acceptance model based on marketing and advertising concepts (Bhattacherjee, 2001), and SERVQUAL (Parasuraman, Berry, & Zeithaml, 1988) for service quality. These models have been extensively used to predict and evaluate online retail shopping and continued acceptance of ICTs. In addition, varieties of integrated models have been developed to measure the success of information systems, ICT, and Internet adoption and diffusion. Currently, many of these models are being tested in the context of mobile technology (primarily mobile phone services).

The integration models mentioned above have been empirically tested in the e-commerce area. The models have been authenticated and proven to be extremely useful in predicting behavior of users of ICT and e-commerce. In the case of m-commerce, the results have been slightly inconsistent. Primarily these inconsistencies have been found because of the differing market maturity levels or the usage pattern of mobile devices. For example, in a South Korean study where mobile phones have been in use for quite some time, the results of testing an integrative m-commerce adoption model yielded different results for actual use than in a similar study conducted in Thailand where mobiles devices were introduced much later in the market. South Koreans were not influenced much by advertising, unlike Thai people in the initial adoption phase of m-commerce. Conversely, Thai people were not influenced by word-of-mouth to the extent South Koreans were influenced in the initial adoption (Thanarithiporn, 2005). According to Thanarithiporn

(2005), this is due to the fact South Koreans are at a more advanced level of adoption for ICTs. Furthermore, Thanarithiporn (2005) found that, unlike in South Korea where content availability had no influence in the continued use of mobile phones, it had a strong influence in Thailand on mobile usage rate. Also, in both countries self-efficacy had no influence one way or the other in the initial adoption of the mobile phone.

Key Factors that Affect the Adoption and Diffusion of E-Commerce and M-Commerce

As expected, many factors influence the rate of adoption and diffusion of technological innovations. We reviewed the extant literature, as outlined above, to identify those factors. In particular, we were interested in a set of factors that have significant influence in the adoption and diffusion of both e-commerce and m-commerce. These include: (a) hardware requirement, (b) software requirement, (c) connection or accessibility, and (d) content. In the following paragraphs, we will outline how these factors have influenced the development of e-commerce, and are currently influencing the adoption and diffusion of m-commerce.

Hardware Requirement

E-Commerce

Computer users were used to the QWERTY keyboard (of typewriters), thus they easily adapted to the standardized desktop of the first personal computers (PCs) in the 1980s. The development of graphical user interface (GUI), mice, and various other multimedia-related accessories has made PCs and variations thereof easy to use. With the introduction of open architecture, the adoption and diffusion of PCs proliferated. The introduction of the Internet to the common public, and the introduction of the GUI browser immediately thereafter, allowed PC users to quickly adopt the Web browsers and demand applications in a hurry. The limitation of hardware at the user level was only restricted by the inherent rendering capability of a model based on the processors, configuration, and accessories that supported them. Since the Web and e-commerce server technologies that serve Internet documents or Web pages are also based on open architecture, limitations were similar to that of desktops.

M-Commerce

The hardware used for mobile devices are complex. The evolution of the hardware technology used in mobile devices is diverse because of the diversity in fundamental architecture. These architectures are based on diverse technology standards such as TDMA, CDMA, GPRS, GSM, CDMA/2000, WCDMA, and i-mode. In addition, these architectures have gone through multiple generations of technology such as 1G (first generation – analog technology); 2G (second generation – digital technology, including 2.5G and 2.75G); and 3G, to meet the demands of customers in terms of bandwidth speed, network capabilities, application base, and corresponding price structures. The lack of uniform global standards and varied sizes and user interfaces to operate the devices has further disrupted the smoother adoption process. While the U.S. still suffers from a lack of uniform standard, Europe is moving towards uniformity through some variation of TDMA technology, and China is modifying CDMA technology to develop its own standard. Other countries are currently working towards a uniform standard based on a variation of base TDMA or CDMA technology (Keen & Mackintosh, 2001).

The innovation in the changing standards, devices, applications, and cultural temperament have constantly maintained a turbulent environment in the adoption and diffusion of commerce through mobile devices. For example, if the device is WAP-enabled, then Web services can be delivered using standardized WML, CHTML, or J2ME development tools. But the WAP enabling has not given scale advantages because hardware standards have not converged, at least not in the U.S. where consumers use a multitude of devices such as Palm, different Web-enabled phones, and different pocket phones.

Software Requirement

E-Commerce

The standardization and open architecture of PCs, along with the high degree of penetration of PCs in the office and home environment, allowed for standardization of client devices. This allowed for the development of text browsers, and subsequently the development of the graphical interface through Web browsers. Apples, PCs, and other UNIX-based workstations were able to use the device-independent Web browsers, thus leading to rapid adoption and expansion in the usage of Web browsers. The low price of earlier browsers such as Mosaic and Netscape, and the distribution of Internet Explorer with the Windows Operating System by Microsoft allowed the diffusion of the browsing capability in almost every client in the market.

Standardized browser software and interface, along with market dominant operating systems such as the Windows family of desktop operating systems and server platforms, facilitated the exponential growth of Internet users and applications. The availability, integration, and interoperability of application development tools, and the reliance on open systems concept and architecture, fueled further changes in the interactivity of the Web and indirectly boosted the commerce on the Web. The development of hardware-independent Java (by Sun Microsystems) and similarly featured tools allowed growth in the interactivity of the Web and application integration both at the front end and backend of the Web. The interoperability of Web applications to communicate with a wide variety of organizational systems initiated a concern for security of the data while in transit and storage. In the early stages of e-commerce, major credit card companies did not trust the methodologies that were used, although they allowed the transactions. Beginning in 1999, they started protecting the online customers just as they protected off-line customers (namely, a customer is only responsible for $50 if she reports the card stolen within 24 hours). The technology companies and financial service organizations collaboratively created and standardized methodologies for online secure transactions, and originated the concept of third-party certification of authority. This certification practice further strengthened the security of online commerce and established a strong basis for consumers to trust and online commerce to grow.

M-Commerce

Software for mobile technologies is dependent on the technology standard used and type of applications suitable for the mobile device. In most nations, like in the U.S., the use of mobile devices started with the use of analog cellular phones. These required proprietary software and proprietary networks. The digitization of handheld devices started with personal digital assistants (PDAs) for personal information management. The transformation of the PDA as a digital communication tool was made possible by private networks, operating systems, and applications developed by companies such as Palm. However, as Microsoft's Windows CE (Compact Edition) and BlackBerry started offering e-mail, information management tools, and Web surfing using micro-browsers, the growth in the use of handheld devices for Web applications started growing. The handheld industry responded with a variety of applications and made WAP a standard for applications development.

Concurrently, the telecom industry brought out digital phones and devices that could offer voice, personal information management (PIM), and data applications. However, until now, operating systems, servers, and Web applications are not standardized in the handheld market. The diversity of server software and client operating systems, and the availability of applications have not made these devices interoperable. In addition, with each player offering its own network and original content or converted content (i.e., content originally developed for the desktop computers), the interest in commerce using mobile devices has not been too enthusiastic. Furthermore, the lack of common security standards has made mobile commerce adoption very slow.

Connection or Access

E-Commerce

In the United States, where telephone wire lines have been in existence for over 100 years, it was natural for the telecom companies to focus on offering Internet connectivity through the existing telephone network. In the early stages of pubic offering of the Internet, it was easy for people to adopt the Internet using their modem from a private network. As the Internet evolved into the World Wide Web, and innovation brought faster modems to the market, more Internet service providers (ISPs) started providing ramps to the Internet. When the Windows98 Operating System with its integrated Internet Explorer was introduced to the marketplace, the Internet adoption was growing in triple digits per year. The major infrastructural components were already in place. The telecom sector invested heavily into building the bandwidth and router network to meet the insatiable demand for Web surfing. Worldwide Internet adoption and use was growing exponentially. The ICT industry responded with innovative technologies, software and services using standardized PCs, modems, support for (Internet protocol suite) TCP/IP protocol of Internet, and highly competitive pricing. The e-tail industry subsequently started growing rapidly, and the financial service industry introduced innovative products and services while collaboratively designing secure electronic payment mechanisms with ICT industry players.

The drop in pricing, availability of bandwidth, security, and quality of products and services bolstered the commerce activity on the Internet until the 'dot.com bust' of May 2000. Although the bust slowed the growth rate of e-commerce, in reality e-commerce continuously grew despite the bust. Support for e-commerce from the U.S. government to fuel the e-commerce growth through moratorium on taxes by two administrations considerably helped the diffusion of e-commerce. The concern about the security in e-commerce shown by laggards was eased by a variety of security and encryption tools, and the creation of the certification of authority concept by strong security services offered by companies such as Verisign, TRUSTe, and others.

Lately, the demand for highly competitive broadband service availability, and the availability and delivery of media-rich content, has brought media and entertainment industry to the Web with greater force. These technological advances in the e-commerce sector have received increased attention, thus ensuring a strong global growth rate in e-commerce.

M-Commerce

In the mobile arena, customers may have been using analog cellular phones (1G) for a long of time. During the era of analog cellular phones, the common mobile commerce activity was the downloading of ring tones. This type of commerce activity is still quite prevalent in developing nations. In addition to this type of commerce, other types of commerce conducted using these devices are the same as the ones that can be performed using a standard desk phone, such as ordering tickets for an event, ordering catalog items, and similar tasks.

With the introduction of digital devices (2G), mobile phones quite suddenly have become the lifeline for many transactions, such as e-mail, voicemail, and text messaging. With 2.5G, 2.75G, and now with 3G devices, more varied and complex applications such as photo transfers, interactive games, and videos have become the norm. The capabilities of these devices are determined by technical ability of the devices and the support of terrestrial tower structures by the vendors offering these services. In addition, the content availability and their desirability by the customers also determine the adoption of such services. The technology, standards, and competition have left U.S. vendors in the distance in rolling out new technology and services. While Asia's (South Korea, Japan, and China) mobile penetration growth is three times that of the United States, Europe is closely behind Asia, with England (87%) and Finland (75%) achieving very high penetration rates (Shim, 2005). In the U.S., the major players in the telecom industry are collaborating to achieve the 3G-standard Universal Mobile Telecommunication System (UMTS) to provide penetration and support rollout of new technology and services. Several countries including South Korea were planning to offer a more advanced technology called the Digital multimedia broadband (DMB) or wire broadband (WiBro) by the end of 2006 (Shim, 2005). According to Shim (2005), it will take a while to obtain DMB cellular phone services in the U.S., since technical standards and logistical barriers will have to be overcome first.

The private networks built by the wireless service providers through the customized devices will determine the access and speed available in the future in the United States. The investment in the network, along with the rollout of new technology and methods used to price the services, will be strong factors in building the capacity. Government policies are also vital in this respect. According to Shim (2005), the government commitment and push for IT strategy and long-term goals are among the most important factors to advance a country's cellular mobile business, particularly for less-developed countries.

Content

E-Commerce

Identifying the most preferred method for delivery of any content has always been a thorny issue. In electronic commerce, the complete digital conversion of all media into technology mandated by the FCC by 2008 would be much easier (FCC, 2006). Voice, as well as radio and television signals, will be broadcast digitally. The Internet has built capacity to deliver rich media content at high speed using the fiber network in the U.S. The convergence of devices such as TV monitors and PC monitors has already brought down the prices for such devices due to scale effects. The stumbling blocks to achieve a greater level of broadband adoption (from the current 53% in the U.S.) are pricing and quality of content (Pew, 2006). In e-commerce, content can be provided by anyone using standardized development tools and can be served on the standardized server software since most desktops can handle all the content delivered through the Web. The diffusion of such innovations is constrained by the pricing and the investment made by consumers at the client level. The industry has converged in standardizing hardware, software, and protocols. Globally as well as in the U.S., there is a clear trend to make the technology affordable throughout the world through the open systems concept. This has helped tremendously, especially in developing countries, in the adoption and diffusion of the Internet and generalized applications.

M-Commerce

In mobile commerce, the content such as data, text, audio, video, and video streaming can be delivered through the devices provided by service providers through their network infrastructure. As the service providers rollout new network technologies with greater capabilities to adapt to the new generation of hardware and software technologies, consumers can expect more media-rich content. Any content that is available in the e-commerce world will be specially modified for mobile delivery using specific development tools for WAP-enabled devices such as WML, CHTML, and J2ME.

Depending on the type of device, the content will have to be delivered in device-specific configuration—for example, the content has to be delivered differently to a PocketPC, WAP-enabled mobile phone, and WAP-enabled PDAs. This type of dynamic configuration in the content delivery requires investment from service providers and/or value-added intermediaries. The special intermediaries provide enormous value-added services in converting the e-commerce content for different mobile devices and become consolidators of content and applications and essentially become data portals for mobile devices. The diversity of devices available in the market will require a significant amount of investments in the U.S. to offer it nationwide, unless it focuses only on high-population density regions to maximize returns.

CONCLUSION

Based on the foregoing discussion, we can say that the introduction of e-commerce has been comparatively smoother than m-commerce. The development of the hardware capability (from PC to GUI to other multimedia-related accessories such as printers, camera, etc.), the software capability (such as browsers, open operating systems, payment schemes, secure systems, etc.), better accessibility (such as phone lines, cables, etc.), and more varied content (such as voice, radio, and television signals) ensured a fast adoption and diffusion of e-commerce throughout the world.

It is true that m-commerce also enjoys many advantages similar to e-commerce. For example, the mobile phone—the principal mode of m-commerce—is witnessing a spectacular growth throughout the world. Unfortunately, unlike e-commerce, m-commerce does not enjoy an open architecture that can accommodate varied standards in hardware, software, connection technology, and the content. Several countries (such as Japan and South Korea) are further ahead of the U.S. in solving this issue of incompatible technologies. It is heartening to see a sincere effort in many countries, including the U.S., to achieve convergence in technologies so that m-commerce is able to grow true to its full potential.

REFERENCES

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

APRG. (2006). Retrieved from http://www.aprg.com

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly, 25*(3), 351-370.

Cassidy, J. (2002). *dot.con: The greatest story every sold.* New York: Harper Collins.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly, 13*(2), 319-339.

DEVX. (2006). Retrieved from http://www.devx.com/wireless/Door/11297

eMarketer. (2005). Mobile marketing and m-commerce: Global spending and trends. *eMarketer,* (February 1).

FCC. (2006). Retrieved from http://www.fcc.gov/cgb/consumerfacts/digitaltv.html

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Keen, P., & Mackintosh, R. (2001). *The freedom economy: Gaining the m-commerce edge in the era of the wireless Internet.* Berkeley, CA: Osborne/McGraw-Hill.

Kini, R. B., & Thanarithiporn, S. (2004). M-commerce and e-commerce in Thailand—A value space analysis. *International Journal of Mobile Communications, 2*(1), 22-37.

Parasuraman, A., Berry, L. L., & Zeithaml, V. A. (1988). SERVQUAL: A multiple-item scale for measuring customer perceptions of service quality. *Journal of Retailing, 64*(1), 12-40.

Pew. (2006). Retrieved from http://www.pewinternet.org

Rogers, E. M. (1995). *Diffusion of innovations.* New York: The Free Press.

Schifter, D. E., & Ajzen, I. (1985). Intention, perceived control, and weight loss: An application of the theory of planned behavior. *Journal of Personality and Social Psychology, 49*(3), 843-851.

Shim, J. P. (2005). Korea's lead in mobile cellular and DMB phone services. *Communications of the Association for Information Systems, 15,* 555-566.

Thanarithiporn, S. (2004). *A modified technology acceptance model for analyzing the determinants affecting initial and post intention to adopt mobile technology in Thailand.* Unpublished dissertation, Bangkok University, Thailand.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

Adoption of M-Commerce Devices by Consumers

Humphry Hung
Hong Kong Polytechnic University, Hong Kong

Vincent Cho
Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

The Internet has undoubtedly introduced a significant wave of changes. The increased electronic transmission capacity and technology further paves a superhighway towards unrestricted communication networks (Chircu & Kauffman, 2000; Cowles, Kiecker, & Little, 2002). It is estimated that by 2007, the total number of Internet users in the world will be over 1.4 billion and the percentage of wireless users is projected to take up about 57% of the vast number (Magura, 2003). Most people anticipate that the next-generation commerce will emerge from traditional commerce to PC-based e-commerce, and eventually to mobile commerce (Ellis-Chadwick, McHardy, & Wiesnhofer, 2000, Miller, 2002, Watson, Pitt, Berthon, & Zinkhan, 2002).

Mobile commerce (m-commerce) is an extension, rather than a complete replacement, of PC-based electronic commerce. It allows users to interact with other users or businesses in a wireless mode, anytime and anywhere (Balasubramanian, Peterson, & Jarvenpaa, 2002; Samuelsson & Dholakia, 2003). It is very likely that PC-based e-commerce will still prevail for a relatively long period of time in spite of the trend that more and more people will choose to adopt m-commerce for their purchases (Miller, 2002).

The focus of our article is on the consumers' adoption of m-commerce devices (MCDs), which are equipment and technologies that facilitate users to make use of m-commerce. MCDs include mobile phones, personal digital assistants (PDA), portable computer notebooks, Bluetooth, WAP, and other facilities that can have access to the wireless networks. We expect that the heading towards a world of mobile networks and wireless devices, which will present a new perspective of time and space, is definitely on its way.

Several basic questions about m-commerce devices will be addressed in this article. First, why should consumers adopt MCDs? What will be the influencing factors for consideration? Are these MCDs easy to use and proven to be useful? Second, how do the MCDs compare with the devices for other types of commerce, such as e-commerce or traditional mail order? Consumers will only adopt MCDs when there are some potential significant advantages when comparing to old devices for other types of commerce. There is still a lack of comprehensive framework within which the adoption of MCDs can be evaluated. Traditional viewpoints regarding this issue, especially those that are based on technology acceptance models, will need to be revisited and revised when consumers are considering such an adoption.

In this article, we propose a framework for identifying the various influencing factors of the adoption of MCD, as well as the antecedents of these influencing factors. Because of the need of the standardization of the application, interface, and inter-connectivity of all hardware and software relevant to the adoption and usage of MCDs, our proposed framework will have some global implications (Zwass, 1996). Our conceptual framework can, therefore, make significant contributions to a more in-depth understanding in the spread and acceptability of m-commerce through knowing why and how relevant MCDs are adopted.

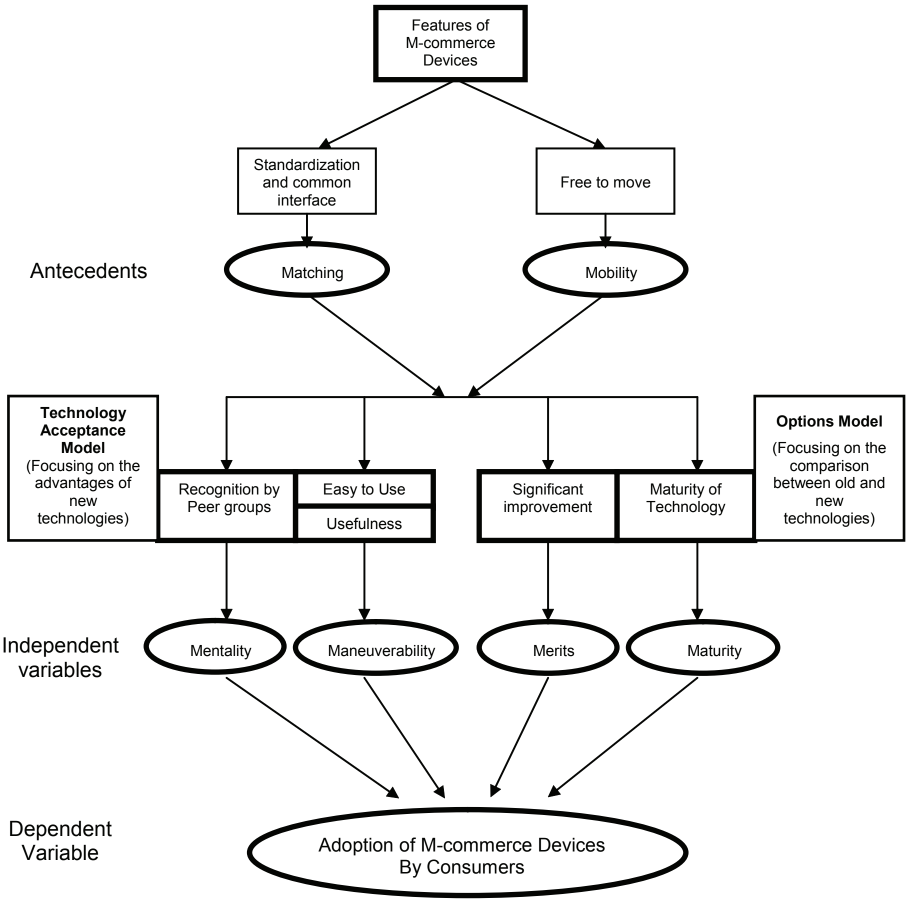
While using technology acceptance models (TAMs) as our primary reference, we also incorporate the important implications of an options model into our basic framework of analyzing consumers' adoption of MCDs. Based on our theoretical framework, we identify four influencing factors—merits, maturity, maneuverability, and mentality—which we consider to be relevant to the decision of consumers in adopting MCDs. We also identify two generic antecedents of these influencing factors—mobility and matching. We plan to investigate the extent of influence of these influencing factors and their antecedents, which will affect consumers' adoption decisions of MCDs. Figure 1 is a graphical representation of our conceptual model of the adoption of MCDs by consumers.

INFLUENCING FACTORS BASED ON TECHNOLOGY ACCEPTANCE MODEL

The technology acceptance model is an information systems theory that models how users come to accept and use a new technology, with reference to two major considerations, perceived usefulness and perceived ease of use (Venkatesh & Davis, 2000). The former is about the degree to which a person believes that using a particular system will make

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. A conceptual model of the adoption of m-commerce devices


Features of M-commerce Devices
Standardization and common interface
Free to move
Antecedents
Matching
Mobility
Technology Acceptance Model (Focusing on the advantages of new technologies)
Recognition by Peer groups
Easy to Use
Usefulness
Significant improvement
Maturity of Technology
Options Model (Focusing on the comparison between old and new technologies)
Independent variables
Mentality
Maneuverability
Merits
Maturity
Dependent Variable
Adoption of M-commerce Devices By Consumers

his or her life easier, for instance, by enhancing his or her job performance or reducing the workload, while the latter is the degree to which a person believes that it is not difficult to actually use a particular system (Venkatesh & Davis, 2000).

With reference to TAM, we consider whether the adoption of MCDs will bring advantages to consumers. We identify two Ms, maneuverability and mentality, for relating the acceptability of MCDs to users.

The first influencing factor, maneuverability, is related to the perceived usefulness in the adoption of MCDs and the degree to which a person can make the best use of such MCDs. Consumers will tend to adopt devices that are user friendly and do not require some intensive training of adoption (Prasanna et al., 1994).

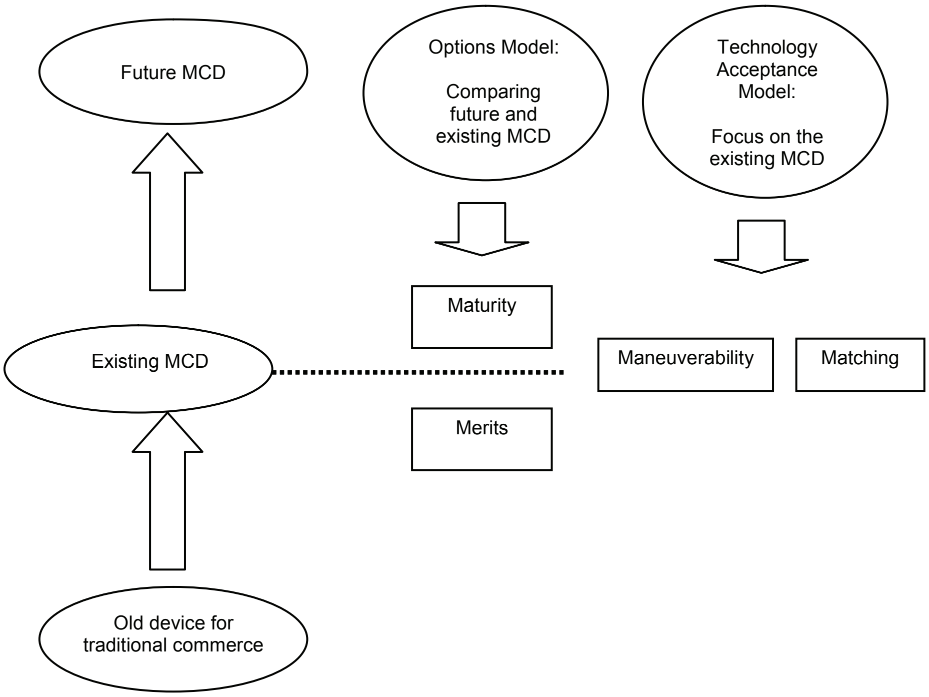
The second influencing factor, mentality, is concerned with the match between the new technology and consumers' own mindsets, as well as the appropriate recognition of their peer groups (Bessen, 1999; Venkatesh & Davis, 2000). General acceptance by the consumers, especially by their peer groups, will be very important to consumers when they consider using MCDs for matching the devices of other people.

INFLUENCING FACTORS BASED ON OPTIONS MODEL

While mainstream literature on the adoption of new technologies is primarily based on the technology acceptance model, we consider that, in the context of m-commerce, we also need to think about some other aspects.

The options model demonstrates that a new technology with a moderate expected improvement in performance can experience substantial delays in adoption and price distortions even in a competitive market (Bessen, 1999; Sheasley, 2000). Rather than adopting a new technology that demonstrates only marginal improvement, consumers have the option of not adopting until the new technology, in terms of performance, is substantially better than the old technology. Consumers contemplating the adoption of a new technology are, of course, aware of the possibility of

Figure 2. A diagrammatic representation of the role of the four influencing factors of MCDs


Future MCD
Options Model:
Comparing future and existing MCD
Technology Acceptance Model:
Focus on the existing MCD
Maturity
Existing MCD
Maneuverability
Matching
Merits
Old device for traditional commerce

sequential improvement. They consider not only the current technical level of the new technology, but also their expectations of possible upgrades and changes in the future of the new technology (Sheasley, 2000).

With regards to the options model, we consider the comparison between MCDs and devices for other types of commerce, and in particular, the comparative advantages of MCDs to consumers. Based on the options model, we identify two Ms, merits and maturity, in relation to the comparison.

We identify the third influencing factor, merits, which is about the degree to which a buyer believes that the MCD can provide significant improvement in the purchase process. Handheld mobile devices, such as PDAs and other enhanced alphanumeric communicators, have supplemented mobile telephones, thus expanding the range of MCDs available for m-commerce transactions. With the abilities to be connected to digital communication networks, MCDs are considered to be in possession of important comparative advantage of mobility.

The fourth influencing factor, maturity, is the possibility that the technology of the MCD is mature enough so that there will not be any possible significant improvements at a later stage. While academic researchers and business practitioners recognize that the electronic market will penetrate and replace a traditional type of commerce, there are still some reservations that will likely cause the early adopters of new technologies some problems in terms of the obsolescence of devices (Samuelsson & Dholakia, 2003). Most consumers will prefer adopting MCDs with more mature technologies so that there is no need for a high level of subsequent upgrading of devices.

In essence, the option model focuses on the comparison between existing and old MCDs, while TAM places emphasis on the generic attributes and utility of MCDs. Figure 2 shows the inter-relationship among the four influencing factors. Based on the four factors that we have identified, we propose the followings:

Proposition 1: Maneuverability, mentality, merits, and maturity are the influencing factors when consumers consider adopting MCDs for purchases.

GENERIC ATTRIBUTES OF MCD

In addition to the identification of the influencing factors of the adoption of MCDs, we also consider their antecedents, which are related to the very basic and essential characteristics of MCDs.

We start our analysis by considering two generic attributes of MCDs, mobility and matching. Mobility is the most fundamental aspect of m-commerce because the name m-commerce arises from the mobile nature of the wireless environment that supports mobile electronic transactions (Coursaris, Hassanein, & Head, 2003). Mobile wireless devices, such as mobile phones, PDAs, and portable computer notebooks, can have the ability to help users gain access to the Internet. Based on these wireless devices, m-commerce is a natural extension of e-commerce but can provide some additional advantages of mobility for consumers. Mobility is a major prerequisite for the adoption of MCDs. It is an antecedent of the influencing factors of the adoption of MCDs because people will consider adopting a wireless connection because it can allow significant improvement compared with traditional device (i.e., merits), and is perceived to be useful and convenient (i.e., maneuverability).

Matching describes the need for the standardized and common interface of MCDs (Coursaris et al., 2003). The unique characteristic of m-commerce very often requires both ends of this new type of commerce to have a common interface. M-commerce applications have the challenging task of discovering services in a dynamically changing environment. Effective mechanisms need to be in place for the interface between various types of MCDs. Matching is an important antecedent of the influencing factors of consumers' adoption of MCDs because the need for standardization (i.e., matching) is important for m-commerce technology, which allows for the connection of MCDs with the wireless networks and the connections among different MCDs. This standardized interface (i.e., matching) also reflects that the MCD is mature (i.e., maturity). Moreover, the standardized interface (matching) will also help to promote the universal acceptance of MCDs by people (i.e., mentality). Based on these arguments, we develop the second proposition:

Proposition 2: The generic attributes of m-commerce, mobility and matching, are the antecedents of the influencing factors when consumers adopt MCDs for purchases.

RESEARCH IMPLICATIONS

Based on our conceptual framework, we identify the various influencing factors (i.e., 4 Ms) which can affect consumers' decisions about the adoption of MCDs in their purchases. It is possible to collect data on whether consumers will consider the adoption of MCDs, and at the same time, researchers can also investigate the reasons why they adopt or do not adopt MCDs, in terms of timing, opportunities, changing trends, and applications.

In our conceptual framework, the dependent variable is the intention of consumers to adopt MCDs. We identify four Ms as the primary influencing factors of the adoption of new technologies in m-commerce (maneuverability, mentality, merits, and maturity). These are independent variables in our framework. We also identify the antecedents of these influencing factors, mobility and matching.

First, maneuverability will be measured by the usability of the MCD. Mentality can be evaluated by the perceived peer groups' acceptance of MCDs. Merits can be measured by the comparative advantages of the MCD in relation to the old devices for other types of commerce. Maturity can be assessed by the perception that the relevant MCD can or cannot be upgraded. The first antecedent, mobility, can be measured by the extent of access to wireless networks. Matching can be measured by the degree that MCDs can be compatible with each other.

In addition to the primary independent variables, we suggest measuring some important control or moderating variables, such as price, and completing a demographic profile such as sex, age, and education levels, as well as occupations and incomes of consumers.

CONCLUSION AND IMPLICATIONS

In our proposed model, we are exploring new insights and new adoption behavior in the ubiquitous world of m-commerce, which we believe is not yet fully understood by most marketers and scholars (Stevens & McElhill, 2000; Struss, El-Ansary, & Frost, 2003). Our proposed model will be of interest to academics in the IT field, who may be keen to know how they can perform further relevant research in m-commerce.

Our proposed framework represents a theory-driven examination of the adoption of MCDs by consumers in their purchase processes. The powerful tool of m-commerce can allow for faster and easier response to market demand, and at the same time consumers can obtain relevant information as well as purchasing goods and services at any time and anywhere as they prefer.

It is expected that our proposed framework can provide important guidelines for pointing the way towards some relevant research on the significance of the adoption of MCDs. Our conceptual framework contributes to literature by ascertaining the most significant independent variables from among all those key variables that we have identified based on our literature review, which can determine which and how new technologies are likely to be adopted in m-commerce.

REFERENCES

Balasubramanian, S., Peterson, R.A., & Jarvenpaa, S.L. (2002). Exploring the implications of m-commerce for markets and marketing. *Academy of Marketing Science Journal, 30*(4), 348-361.

Bessen, J. (1999). *Real options and the adoption of new technologies.* Retrieved from http://www.researchoninnovation.org/online.htm#realopt

Bisbal, J., Lawless, D., Wu, B., & Grimson, J. (1999). Legacy information system migration: A brief review of problems, solutions and research issues. *IEEE Software, 16,* 103-111.

Chircu, A., & Kauffman, R. (2000). Reintermediation strategies in business-to-business electronic commerce. *International Journal of Electronic Commerce*, *4*(4), 7-42.

Coursaris, C., Hassanein, K., & Head, M. (2003). M-commerce in Canada: An interaction framework for wireless privacy. *Canadian Journal of Administrative Sciences, 20*(1), 54-73.

Cowles, D.L., Kiecker, P., & Little, M.W. (2002). Using key informant insights as a foundation for e-retailing theory development. *Journal of Business Research, 55,* 629-636.

Ellis-Chadwick, F., McHardy, P., & Wiesnhofer, H. (2000). Online customer relationships in the European financial services sector: A cross-country investigation. *Journal of Financial Services Marketing*, *6*(4), 333-345.

Magura, B. (2003). *What hooks m-commerce customers? MIT Sloan Management Review, 44*(3), 9-10.

Miller, A.I. (2002). *Einstein, Picasso: Space, time, and the beauty that causes havoc.* New York: Basic Books.

Samuelsson, M., & Dholakia, N. (2003). Assessing the market potential of network-enabled 3G m-business services. In S. Nansi (Ed.), *Wireless communications and mobile commerce.* Hershey, PA: Idea Group Publishing.

Sheasley, W.D. (2000). Taking an options approach to new technology development. *Research Technology Management, 43*(6), 37-43.

Stevens, G.R., & McElhill, F. (2000). A qualitative study and model of the use of e-mail in organizations. *Electronic Networking Applications and Policy, 10*(4), 271-283.

Struss, J., El-Ansary, A., & Frost, R. (2003). *E-marketing* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Venkatesh, V., & Davis, F.D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

Watson, R.T., Pitt, L.F., Berthon, P., & Zinkhan, G.M. (2002). U-commerce: Extending the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 329-343.

Zwass, V. (1996). Electronic commerce: Structures and issues. *International Journal of Electronic Commerce, 1*(1), 3-23.

KEY TERMS

Maneuverability: The perceived usefulness in the adoption of MCDs and the degree to which a person can make the best use of such MCDs; one of the influencing factors when consumers consider adopting MCDs for purchases.

Matching: The need for the standardized and common interface of MCDs.

Maturity: The possibility that the technology of the MCD is mature enough so that there will not be any possible significant improvements at a later stage; one of the influencing factors when consumers consider adopting MCDs for purchases.

MCD: M-commerce device.

Mentality: The match between the new technology and consumers' own mindsets, as well as the appropriate recognition of their peer groups; one of the influencing factors when consumers consider adopting MCDs for purchases.

Merits: The degree to which a buyer believes that MCDs can provide significant improvement in the purchase process; one of the influencing factors when consumers consider adopting MCDs for purchases.

Options Model: A model that proposes that consumers have the option of not adopting until the new technology, in terms of performance, is substantially better than the old technology, and as a result of such options, a new technology with a moderate expected improvement in performance can experience substantial delays in adoption and price distortions even in a competitive market.

Technology Acceptance Model (TAM): An information systems theory that models how users come to accept and use a new technology, with reference to two major considerations, perceived usefulness and perceived ease of use.

Advanced Resource Discovery Protocol for Semantic-Enabled M-Commerce

Michele Ruta
Politecnico di Bari, Italy

Tommaso Di Noia
Politecnico di Bari, Italy

Eugenio Di Sciascio
Politecnico di Bari, Italy

Francesco Maria Donini
Università della Tuscia, Italy

Giacomo Piscitelli
Politecnico di Bari, Italy

INTRODUCTION

New mobile architectures allow for stable networked links from almost everywhere, and more and more people make use of information resources for work and business purposes on mobile systems. Although technological improvements in the standardization processes proceed rapidly, many challenges, mostly aimed at the deployment of value-added services on mobile platforms, are still unsolved. In particular the evolution of wireless-enabled handheld devices and their capillary diffusion have increased the need for more sophisticated service discovery protocols (SDPs).

Here we present an approach, which improves Bluetooth SDP, to provide m-commerce resources to the users within a piconet, extending the basic service discovery with semantic capabilities. In particular we exploit and enhance the SDP in order to identify generic resources rather than only services.

We have integrated a "semantic layer" within the application level of the standard Bluetooth stack in order to enable a simple interchange of semantically annotated information between a mobile client performing a query and a server exposing available resources.

We adopt a simple piconet configuration where a stable networked zone server, equipped with a Bluetooth interface, collects requests from mobile clients and hosts a semantic facilitator to match requests with available resources. Both requests and resources are expressed as semantically annotated descriptions, so that a *semantic distance* can be computed as part of the ranking function, to choose the most promising resources for a given request.

STATE OF THE ART

Usually, resource discovery protocols involve a requester, a lookup or directory server and finally a resource provider. Most common SDPs, as service location protocol (SLP), Jini, UPnP (Universal Plug aNd Play), Salutation or UDDI (universal description discovery and integration), include registration and lookup of resources as well as matching mechanisms (Barbeau, 2000).

All these systems generally work in a similar manner. Basically a client issues a query to a directory server or to a specific resource provider. The request may explicitly contain a resource name with one or more attributes. The lookup server—or directly the resource provider—attempts to match the query pattern with resource descriptions stored in its database, then it replies to the client with discovered resources identification and location (Liu, Zhang, Li, Zhu, & Zhang, 2002).

These discovery architectures are based on some common assumptions about network infrastructure under the application layer in the protocol stack. In particular, current SDPs usually require a continuous and robust network connectivity, which may not be the case in wireless contexts, and especially in the ad-hoc ones. In fact in such environments, network consistence varies continuously and temporary disconnections occur frequently, so bringing to a substantial decrease traditional SDP performances (Chakraborty, Perich, Avancha, & Joshi, 2001).

Actually, there are several issues that restrain the expansion of advanced wireless applications, among them, the variability of scenarios. An ad-hoc environment is based on short-range, low power technologies like Bluetooth

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

(Bluetooth, 1999), which grant the peer-to-peer interaction among hosts. In such a mobile infrastructure there could be one or more devices providing and using resources but, as a MANET is a very unpredictable environment, a flexible resource search system is needed to overcome difficulties due to the host mobility. Furthermore, existing mobile resource discovery methods use simple string-matching, which is largely inefficient in advanced scenarios as the ones related to electronic commerce. In fact, in these cases there is the need to submit articulate requests to the system to obtain adequate responses (Chakraborty & Chen, 2000).

With specific reference to the SDP in the Bluetooth stack, it is based on a 128-bit universally unique identifier (UUID); each numeric ID is associated to a single service class. In other words, Bluetooth SDP is code-based and consequently it can handle only exact matches. Yet, if we want to search and retrieve resources whose description cannot be classified within a rigid schema (e.g., the description of goods in a shopping mall), a more powerful discovery architecture is needed (Avancha, Joshi, & Finin, 2002). SDP should be able to cope with non-exact matches (Chakraborty & Chen, 2000), and to provide a ranked list of discovered resources, computing a distance between each retrieved resource and the request after a matchmaking process.

To achieve these goals, we exploit both theoretical approach and technologies of semantic Web vision and adapt them to small ad-hoc networks based on the Bluetooth technology (Ruta, Di Noia, Di Sciascio, Donini, & Piscitelli, 2005).

In a semantic-enabled Web—what is known as the semantic Web vision—each available resource should be annotated using RDF (RDF Primer, 2004), with respect to an OWL ontology (Antoniou & van Harmelen, 2003). There is a close relation between the OWL-DL subset of OWL and description logics (DLs) (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2002) semantics, which allows the use of DLs-based reasoners in order to infer new information from the one available in the annotation itself.

In the rest of the article we will refer to DIG (Bechhofer, 2003) instead of OWL-DL because it is less verbose and more compact: a good characteristic in an ad-hoc scenario. DIG can be seen as a syntactic variant of OWL-DL.

THE PROPOSED APPROACH

In what follows we outline our framework and we sketch the rationale behind it. We adopt a mobile commerce context as reference scenario.

In our mobile environment, a user contacts via Bluetooth a zone resource provider (from now on *hotspot*) and submits her semantically annotated request in DIG formalism. We assume the zone server—which classifies resource contents by means of an OWL ontology—has previously identified shopping malls willing to promote their goods and it has already collected semantically annotated descriptions of goods. Each resource in the m-marketplace owns an URI and exposes its OWL description.

The *hotspot* is endowed with a *MatchMaker* [in our system we adapt the MAMAS-tng reasoner (Di Noia, Di Sciascio, Donini, & Mongiello, 2004)], which carries out the matchmaking process between each compatible offered resource and the requested one measuring a "semantic distance." The provided result is a list of discovered resources matching the user demand, ranked according to their degree of correspondence to the demand itself.

By integrating a semantic layer within the OSI Bluetooth stack at service discovery level, the management of both syntactic and semantic discovery of resources becomes possible. Hence, the Bluetooth standard is enriched by new functionalities, which allow to maintain a backward compatibility (handheld device connectivity), but also to add the support to matchmaking of semantically annotated resources. To implement matchmaking and ontology support features, we have introduced a *semantic service discovery* functionality into the stack, slightly modifying the existing Bluetooth discovery protocol.

Recall that SDP uses a simple request/response method for data exchange between SDP client and SDP server (Gryazin, 2002). We associated unused classes of 128-bit UUIDs in the original Bluetooth standard to mark each specific ontology and we call this identifier *OUUID* (*ontology universally unique identifier*). In this way, we can perform a preliminary exclusion of supply descriptions that do not refer to the same ontology of the request (Chakraborty, Perich, Avancha, & Joshi, 2001). With OUUID matching we do not identify a single service, but directly the context of resources we are looking for, which can be seen as a class of similar services. Each resource semantically annotated is stored within the *hotspot* as resource record. A 32-bit identifier is uniquely associated to a semantic resource record within the *hotspot*, which we call *SemanticResourceRecordHandle*. Each resource record contains general information about a single semantic enabled resource and it entirely consists of a list of resource attributes. In addition to the *OUUID* attribute, there are *ResourceName*, *ResourceDescription*, and a variable number of *ResourceUtilityAttr_i* attributes (in our current implementation 2 of them). *ResourceName* is a text string containing a human-readable name for the resource, the second one is a text string including the resource description expressed in DIG formalism and the last ones are numeric values used according to specific applications. In general, they can be associated to context-aware attributes of a resource (Lee & Helal, 2003), as for example its price or the physical distance it has from the *hotspot* (expressed in metres or in terms of needed time to get to the resource). We use them as parameters of the overall *utility function* that computes matchmaking results.

Table 1. List of PDU IDs with corresponding descriptions

PDU ID	Description
0x00	Reserved
0x01	SDP_ErrorResponse
0x02	SDP_ServiceSearchRequest
0x03	SDP_ServiceSearchResponse
0x04	SDP_ServiceAttributeRequest
0x05	SDP_ServiceAttributeResponse
0x06	SDP_ServiceSearchAttributeRequest
0x07	SDP_ServiceSearchAttributeResponse
0x08	SDP_OntologySearchRequest
0x09	SDP_OntologySearchResponse
0x0A	SDP_SemanticServiceSearchRequest
0x0B	SDP_SemanticServiceSearchResponse
0x0C-0xFF	Reserved

In particular, to allow the representation and the identification of a semantic resource description we introduced in the data representation of the original Bluetooth standard two new *data element type descriptor*: OUUID and DIG text string. The first one is associated to the type descriptor value 9 whereas to the second one corresponds the type descriptor value 10 (both reserved in the original standard). We will associate 1, 2, 4 byte as valid size for the first one and 5, 6, 7 for the DIG text string.

Since the communication is referred to the peer layers of the protocol stack, each transaction is represented by one request Protocol Data Unit (PDU) and another PDU as response. If the SDP request needs more than a single PDU (this case is frequent enough if we use semantic service discovery) the SDP server generates a partial response and the SDP client waits for the next part of the complete answer.

By adding two SDP features *SDP_OntologySearch* (request and response) and *SDP_SemanticServiceSearch* (request and response) to the original standard (exploiting not used PDU ID) we inserted together with the original SDP capabilities further semantic-enabled resource search functions (see Table 1).

The transaction between service requester and *hotspot* starts after ad-hoc network creation. When a user becomes a member of a MANET, she is able to ask for a specific service/resource (by submitting a semantic-based description). The generic steps, up to response providing, for a service request are detailed in the following:

1. The user searches for a specific ontology identifier by submitting one or more *OUUID*$_R$ she manages by means of her client application

Table 2. SDP_OntologySearchRequest PDU parameters

PDU ID	parameters
0x08	- ***OntologySearchPattern*** - *ContinuationState*

Table 3. SDP_OntologySearchResponse PDU parameters

PDU ID	parameters
0x09	- ***TotalOntologyCount*** - ***OntologyRetrievedPattern*** - *ContinuationState*

2. The *hotspot* selects OUUIDs matching each *OUUID*$_R$ and replies to the client
3. The user sends a service request (*R*) to the *hotspot*
4. The *hotspot* extracts descriptions of each resource cached within the *hotspot* itself, which is classified with the previously selected *OUUID*$_R$
5. The *hotspot* performs the matchmaking process between *R* and selected resources it shares. Taking into account the matchmaking results, all the resources are ranked with respect to *R*
6. The *hotspot* replies to the user.

It is important to remark that basically all the previous steps are based on the original SDP in Bluetooth. No modifications are made to the original structure of transactions, but simply we differently use the SDP framework. In what follows we outline the structure of the SDP PDUs we added within the original framework to allow semantic resource discovery.

The first one is the *SDP_OntologySearchRequest* PDU. Their parameters are shown in Table 2.

The *OntologySearchPattern* is a data element sequence where each element in the sequence is a OUUID. The sequence must contain at least 1 and at most 12 OUUIDs, as in the original standard. The list of OUUIDs is an ontology search pattern. The *ContinuationState* parameter maintains the same purpose of the original Bluetooth (Bluetooth, 1999).

The *SDP_OntologySearchResponse* PDU is generated by the previous PDU. Their parameters are reported in Table 3.

The *TotalOntologyCount* is an integer containing the number of ontology identifiers matching the requested ontology pattern. Whereas the *OntologyRetrievedPattern* is a data element sequence where each element in the sequence is a OUUID matching at least one sent with the *OntologySearchPattern*. If no OUUID matches the pattern,

Table 4. SDP_SemanticServiceSearchRequest PDU parameters

PDU ID	parameters
0x0A	- ***SemanticResourceDescription*** - ***ContextAwareParam1*** - ***ContextAwareParam2*** - *MaximumResourceRecordCount* - *ContinuationState*

Table 5. SDP_SemanticServiceSearchResponse PDU parameters

PDU ID	parameters
0x0B	- *TotalResourceRecordCount* - *CurrentResourceRecordCount* - ***SemanticResourceRecordHandleList*** - *ContinuationState*

the *TotalOntologyCount* is set to 0 and the *OntologyRetrievedPattern* contains only a specific OUUID able to allow the browsing by the client of all the OUUIDs managed by the *hotspot* (see the following *ontology browsing* mechanism for further details). Hence the pattern sequence contains at least 1 and at most 12 OUUIDs.

The *SDP_SemanticServiceSearchRequest* PDU follows previous PDU. Their parameters are shown in Table 4.

The *SemanticResourceDescription* is a data element text string in DIG formalism representing the resource we are searching for; *ContextAwareParam1* and *ContextAwareParam2* are data element unsigned integers. In our case study, which models an m-marketplace in an airport terminal, we use them respectively to indicate a reference price for the resource and the hour of the scheduled departure of the flight. Since a generic client interacting with a *hotspot* is in its range, using the above PDU parameter she can impose—among others—a proximity criterion in the resource discovery policy.

The *SDP_SemanticServiceSearchResponse* PDU is generated by the previous PDU. Their parameters are reported in Table 5.

The *SemanticResourceRecordHandleList* includes a list of resource record handles. Each of the handles in the list refers to a resource record potentially matching the request. Note that this list of service record handles does not contain header fields, but only the 32-bit record handles. Hence, it does not have the data element format. The list of handles is arranged according to the relevance order of resources, excluding resources not compatible with the request. The other parameters maintain the same purpose of the original Bluetooth (Bluetooth, 1999).

In all the previous cases, the error handling is managed with the same mechanisms and techniques of Bluetooth standard (Bluetooth, 1999).

Notice that each resource retrieval session starts after settling between client and server the same ontology identifier (OUUID).

Nevertheless if a client does not support any ontology or if the supported ontology is not managed by the *hotspot*, it is desirable to discover what kind of merchandise class (and then what OUUIDs) are handled by the zone server without any a priori information about resources. For this purpose we use the *service browsing* feature (Bluetooth, 1999) in a slightly different fashion with respect to the original Bluetooth standard, so calling this mechanism *ontology browsing*. It is based on an attribute shared by all semantic enabled resource classes, the *BrowseSemanticGroupList* attribute which contains a list of OUUIDs. Each of them represents the browse group a resource may be associated with for browsing.

Browse groups are organized in a hierarchical fashion, hence when a client desires to browse a *hotspot* merchandise class, she can create an *ontology search pattern* containing the OUUID that represents the *root browse semantic group*. All resources that may be browsed at the top level are made members of the *root browse semantic group* by having the root browse group OUUID as a value within the *BrowseSemanticGroupList* attribute.

Generally a *hotspot* supports relatively few merchandise classes, hence all of their resources will be placed in the root browse group. However, the resources exposed by a provider may be organised in a browse group hierarchy, by defining additional browse groups below the root browse group.

Having determined the goods category and the corresponding reference ontology, the client can also download a DIG version of it from the *hotspot* as *.jar* file [such a file extension—among other things—also allows a total compatibility with the Connected Limited Device Configuration (CLDC) technology].

Also notice that since the proposed approach is fully compliant with semantic Web technologies, the user exploits the same semantic enabled descriptions she may use in other Semantic Web compliant systems (e.g., in the Web site of a shopping mall). That is, there is no need for different customized resource descriptions and modelling, if the user employs different applications either on the Web or in mobile systems. The syntax and formal semantics of the descriptions is unique with respect to the reference ontology and can be shared among different environments.

In e-commerce scenarios, the match between demand and supply involves not only the description of the good but also data-oriented properties. It would be quite strange to have a commercial transaction without taking into account price, quantity, and availability, among others. The demander usually specifies how much she is willing to pay, how many items she wants to buy, and the delivery date. Hence, the overall match value depends not only on the distance between the (semantic-enabled) description of the demand and of the supply. It has to take into account the description distance with the difference of (the one asked by the demander and the other proposed by the seller), quantity, and delivery date. The overall utility function combines all these values to give a global value representing the match degree.

Also notice that, in m-commerce applications, in addition to "commercial" parameters also context-aware variables should influence matching results. For example, in our airport case study, we consider the price difference but also the physical distance between requester and seller to weigh the match degree. The distance becomes an interesting value since a user has a temporal deadline for shopping: the scheduled hour of her flight. Hence, a resource might be chosen also according to its proximity to the user.

We will express this distance in terms of time to elapse for reaching the shop where a resource is, leaving from the *hotspot* area. In such a manner the *hotspot* will exclude resources not reachable by the user while she is waiting for boarding and it will assign to resources unlikely reachable (farther) a weight smaller than one assigned to easily reachable ones.

The above approach can be further extended to other data-type properties.

The utility function we used depends on:

- p_D : price specified by the demander
- p_O : price specified by the supplier
- t_D : time interval available to the client
- t_O : time to reach the supplier and come back, leaving from the *hotspot* area
- *s_match*: score computed during the semantic matchmaking process, computed through *rankPotential* (Di Noia, Di Sciascio, Donini, & Mongiello, 2004) algorithm.

$$u(s\_match, p_D, p_O, t_D, t_O) = \frac{s\_match}{2} + \frac{\tanh\frac{t_D - t_O}{\beta}}{3} + \frac{(1+\alpha)p_D - p_O}{6(1+\alpha)p_D} \qquad (1)$$

Notice that p_D is weighted by a (1+α) factor. The idea behind this weight is that, usually, the demander is willing to pay up to some more than what she originally specified on condition that she finds the requested item, or something very similar. In the tests we carried out, we find α=0.1 and β=10 are values in accordance with user preferences. These values seem to be in some accordance with experience, but they could be changed according to different specific considerations.

RUNNING EXAMPLE

A simple example can clarify the rationale of our setting. Here we will present a case study analogous to the one presented in Avancha, Joshi, and Finin (2002), and we face it by means of our approach.

Let us suppose a user is in a duty free area of an airport, she is waiting for her flight to come back home and she is equipped with a wireless-enabled PDA. She forgot to buy a present for her beloved little nephew and now she wants to purchase it from one of the airport gift stores.

In particular she is searching for a learning toy strictly suitable for a kid (she dislikes a child toy or a baby toy) and possibly the toy should not have any electric power supply.

Clearly this request is too complex to be expressed by means of standard UUID Bluetooth SDP mechanism. In addition, non-exact matches between resource request and offered ones is highly probable and the on/off matching system provided by the original standard in this case could be largely inefficient.

Hence both the semantic resource request and offered ones can be expressed in a DIG statement exploiting DL semantics and encapsulated in an SDP PDU.

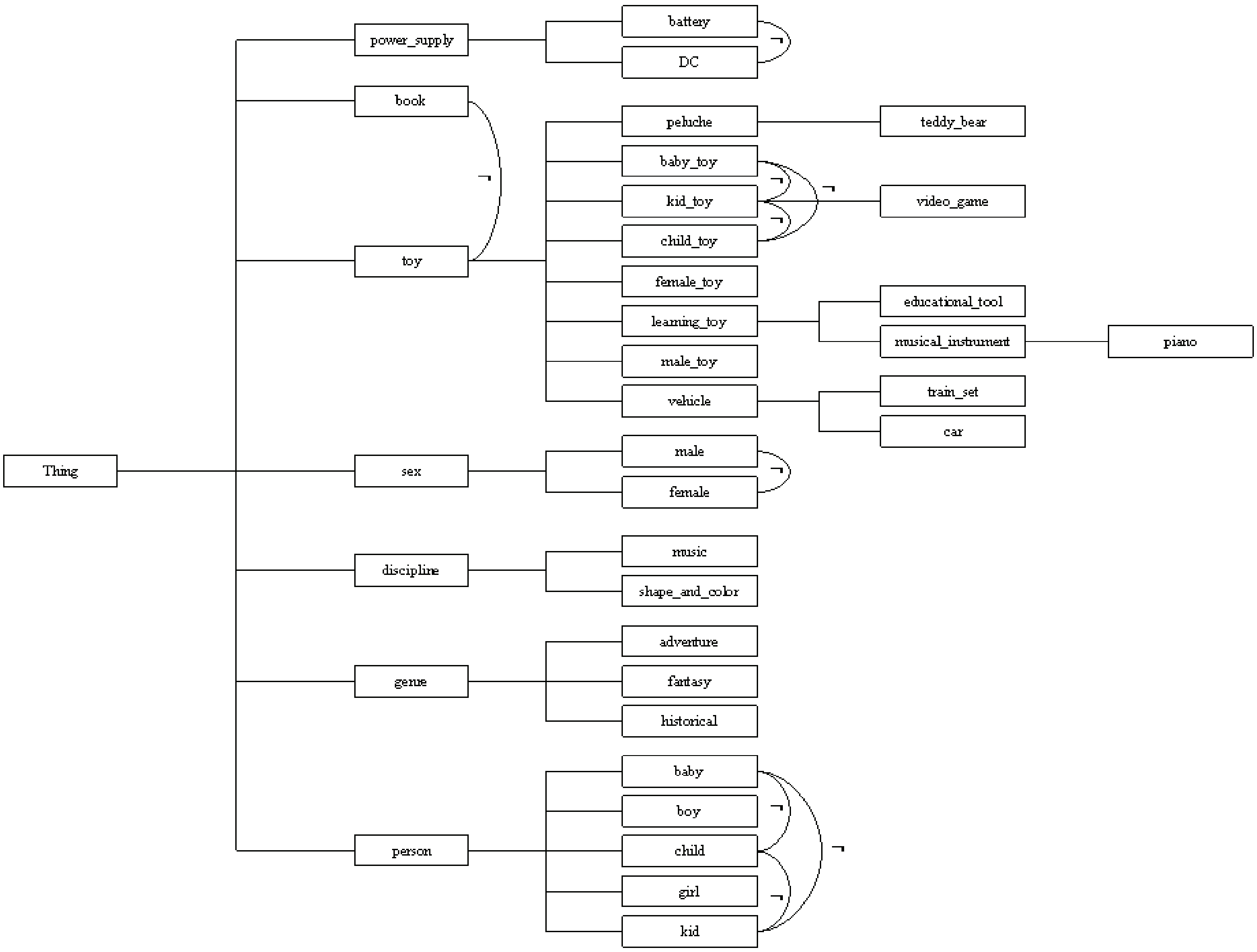
The *hotspot* equipped with MAMAS reasoner collects the request and initially selects supplies expressed by means of the same ontology shared with the requester. Hence a primary selection of suitable resources is performed. In addition, the matchmaker carries out the matchmaking process between each offered resource in the m-marketplace and the requested one measuring a "semantic distance" (Colucci, Di Noia, Di Sciascio, Donini, & Mongiello, 2005). Finally the matchmaking results are ranked and returned to the user.

A subset of the ontology used as a reference in the examples is reported in Figure 1. For the sake of simplicity, only the class hierarchy and disjoint relations are represented.

Let us suppose that after the *hotspot* selects supplies, its knowledge base is populated with the following individuals whose description is represented using DL formalism:

- *Alice_in_wonderland.* Price 20$. 5 min from the *hotspot*:
  book ⊓ ∀has_genre.fantasy
- *Barbie_car.* Price 80$. 10 min from the *hotspot*:
  car ⊓ ∀suggested_for.girl ⊓ ∀has_power_supply.battery
- *classic_guitar.* Price 90$. 17 min from the *hotspot*:
  musical_instrument ⊓ ∀suitable_for.kid ⊓ (£ 0 has_power_supply)
- *shape_order.* Price 40$. 15 min from the *hotspot*:
  educational_tool ⊓ ∀suitable_for.child ⊓ ∀stimulates_to_learn.shape_and_color
- *Playstation.* Price 160$. 28 min from the *hotspot*:
  video_game ⊓ ∀has_power_supply.DC

Figure 1. The simple toy store ontology used as reference in the example


Thing
power_supply
book
toy
sex
discipline
genre
person
battery
DC
peluche
baby_toy
kid_toy
child_toy
female_toy
learning_toy
male_toy
vehicle
teddy_bear
video_game
educational_tool
musical_instrument
piano
train_set
car
male
female
music
shape_and_color
adventure
fantasy
historical
baby
boy
child
girl
kid
¬

- *Winnie_the_pooh.* Price 30$. 15 min from the *hotspot*:
  teddy_bear ⊓ ∀suitable_for.baby

On the other hand, the request ***D*** submitted to the system by the user can be formalized in DL syntax as follows:

learning_toy ⊓ ∀suggested_for.boy ⊓ ∀suitable_for.kid ⊓ (£ 0 has_power_supply)

In addition she imposes a reference price of 200$ (p_D=200) as well as the scheduled departure time as within 30 minutes (t_D=30).

In Table 6 matchmaking results are presented. The second column shows whether each retrieved resource is compatible or not with request *D* and, in case, the *rankPotential* computed result. In the fourth column, matchmaking results are also expressed in a relative form between 0 and 1 to allow a more immediate semantic comparison among requests and different resources and to put in a direct correspondence various rank values.

Finally in the last column results of the overall utility function application are shown.

Notice that the semantic distance of the individual *classic_guitar* from *D* is the smaller one; then the system will recommend to the user this resource first. Hence the ranked list returned by the *hotspot* is a strict indication for the user about best available resources in the airport duty free piconet in order of relevance with respect to the request. Nevertheless

Table 6. Matchmaking results

demand – supply	**compatibility (y/n)**	**score**	**s_match**	**u(·)**
D - Alice_in_wonderland	n	-	-	-
D - Barbie_car	y	7	0.364	0.609
D - classic_guitar	y	3	0.727	0.748
D - shape_order	n	-	-	-
D - Playstation	y	5	0.546	0.378
D - Winnie_the_pooh	n	-	-	-

a user can choose or not a resource according to her personal preferences and her initial purposes.

After having selected the best resource, the server of the chosen virtual shop will receive a connection request from the user PDA with its connection parameters and in this manner the transaction may start. The user can provide her credit card credentials, so that when she reaches the store, her gift will be already packed. This final part of the application is not yet implemented, but it is trivially achievable exploiting the above SDP infrastructure.

CONCLUSION AND FUTURE WORK

In this article we have presented an advanced semantic-enabled resource discovery protocol for m-commerce applications. The proposed approach aims to completely recycle the basic functionalities of the original Bluetooth service discovery protocol by simply adding semantic capabilities to the classic SDP ones and without introducing any change in the regular communication work of the standard. A matchmaking algorithm is used to measure the semantic similarity among demand and resource descriptions.

Future trends of the proposed framework aim to create a more advanced DSS to help a user in a generic m-marketplace. Under investigation is the support to creation of P2P small communities of mobile hosts where goods and resources are advertised and opinions about shopping are exchanged (Avancha, D'Souza, Perich, Joshi, & Yesha, 2003). If a user decides to "open" her shopping trolley sharing information she owns (purchased goods, discounts, opinion about specific vendors or products) the system will insert her in a buyer mobile community where she can exchange information with other users.

Another future activity focuses on strict control of the good advertising. In an m-marketplace, the system will send to various potential buyers best proposals about their interests.

We intend to implement a mechanism to advertise goods or services in a more direct and personalized fashion. From this point of view, an additional feature of the system is oriented to the user profiling extraction and management (Prestes, Carvalho, Paes, Lucena, & Endler, 2004; Ruta, Di Noia, Di Sciascio, Donini, & Piscitelli, 2005; von Hessling, Kleemann, & Sinner, 2004). Without imposing any explicit profile submission to the user, the system could collect her preferences by means of previously submitted requests (Ruta, Di Noia, Di Sciascio, Donini, & Piscitelli, 2005); that is, by means of the "history" of the user in the m-marketplace.

REFERENCES

Antoniou, G., & van Harmelen, F. (2003). Web ontology language: OWL. In *Handbook on Ontologies in Information Systems.*

Avancha, S., D'Souza, P., Perich, F., Joshi, A., & Yesha, Y. (2003). P2P m-commerce in pervasive environments*. ACM SIGecom Exchanges*, *3*(4), 1-9.

Avancha, S., Joshi, A., & Finin, T. (2002). Enhanced service discovery in Bluetooth. *IEEE Computer*, *35*(6), 96-99.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P. (2002). *The description logic handbook*. Cambridge: Cambridge University Press.

Barbeau, M. (2000). Service discovery protocols for ad hoc networking. *Workshop on Ad-hoc Communications (CASCON '00)*.

Bechhofer, S. (2003). *The DIG description logic interface: DIG/1.1*. Retrieved from http://dlweb.man.ac.uk/dig/2003/02/interface.pdf.

Bluetooth specification document. (1999). Retrieved from http://www.bluetooth.com.

Chakraborty, D., & Chen, H. (2000). Service discovery in the future for mobile commerce. *ACM Crossroads*, *7*(2), 18-24.

Chakraborty, D., Perich, F., Avancha, S., & Joshi, A. (2001). Dreggie: Semantic service discovery for m-commerce applications. In *Workshop on Reliable and Secure Applications in Mobile Environment*.

Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F. M., & Mongiello, M. (2005). Concept abduction and contraction for semantic-based discovery of matches and negotiation spaces in an e-marketplace. *Electronic Commerce Research and Applications*, *4*(4), 345-361.

Di Noia, T., Di Sciascio, E., Donini, F. M., & Mongiello, M. (2004). A system for principled matchmaking in an electronic marketplace. *International Journal of Electronic Commerce*, *8*(4), 9-37.

Gryazin, E. (2002). *Service discovery in Bluetooth*. Retrieved from http://www.hpl.hp.com/techreports/2002/HPL-2002-233.pdf.

Lee, C., & Helal, S. (2003). Context attributes: An approach to enable context awareness for service discovery. In *Symposium on Applications and the Internet (SAINT '03)* (pp. 22-30).

Liu, J., Zhang, Q., Li, B., Zhu, W., & Zhang, J. (2002). A unified framework for resource discovery and QoS-aware provider selection in ad hoc networks. *ACM Mobile Computing and Communications Review*, *6*(1), 13-21.

Prestes, R., Carvalho, G., Paes, R., Lucena, C., & Endler, M. (2004). Applying ontologies in open mobile systems. In *Workshop on Building Software for Pervasive Computing OOPSLA'04*.

RDF Primer-W3C Recommendation. (2004, February 10). Retrieved from http://www.w3.org/TR/rdf-primer/

Ruta, M., Di Noia, T., Di Sciascio, E., Donini, F.M., & Piscitelli, G. (2005). Semantic based collaborative P2P in ubiquitous computing. In *IEEE/WIC/ACM International Conference Web Intelligence 2005 (WI '05)* (pp. 143-149).

von Hessling, A., Kleemann, T., & Sinner, A. (2004). Semantic user profiles and their applications in a mobile environment. In *Artificial Intelligence in Mobile Systems 2004*.

KEY TERMS

Description Logics (DLs): A family of logic formalisms for knowledge representation. Basic syntax elements are concept names, role names, and individuals. Intuitively, concepts stand for sets of objects, and roles link objects in different concepts. Individuals are used for special named elements belonging to concepts. Basic elements can be combined using constructors to form concept and role expressions, and each DL has its own distinct set of constructors. DL-based systems are equipped with reasoning services: logical problems whose solution can make explicit knowledge that was implicit in the assertions.

M-Marketplace: Virtual environment where demands and supplies (submitted or offered by users equipped with mobile devices) encounter each other.

Ontology: An explicit and formal description referred to concepts of a specific domain (classes) and to relationships among them (roles or properties).

Piconet: Bluetooth-based short-range wireless personal area network. A Bluetooth piconet can host up to eight mobile devices. More piconets form a *scatternet*.

Service Discovery Protocol (SDP): It identifies the application layer of an OSI protocol stack and manages the automatic detection of devices with joined services.

Semantically Annotated Resource: any kind of good, tangible or intangible (e.g., a document, an image, a product or a service) endowed of a description that refers to a shared ontology.

Semantic Matchmaking: The process of searching the space of possible matches between a request and several resources to find those best matching the request, according to given semantic criteria. It assumes that both the request and the resources are annotated according to a shared ontology.

Anycast-Based Mobility

István Dudás
Budapest University of Technology and Economics, Hungary

László Bokor
Budapest University of Technology and Economics, Hungary

Sándor Imre
Budapest University of Technology and Economics, Hungary

INTRODUCTION

We have entered the new millennium with two great inventions, the Internet and mobile telecommunication, and a remarkable trend of network evolution toward convergence of these two achievements. It is an evident step to combine the advantages of the Internet and the mobile communication methods together in addition to converge the voice and data into a common packet-based and heterogeneous network infrastructure. To provide interworking, the future systems have to be based on a universal and widespread network protocol, such as Internet protocol (IP) which is capable of connecting the various wired and wireless networks (Macker, Park, & Corson, 2001).

However, the current version of IP has problems in mobile wireless networks; the address range is limited, IPv4 is not suitable to efficiently manage mobility, support real-time services, security, and other enhanced features. The next version, IPv6 fixes the problems and also adds many improvements to IPv4, such as extended address space, routing, quality of service, security (IPSec), network autoconfiguration and integrated mobility support (Mobile IPv6).

Today's IP communication is mainly based on unicast (one-to-one) delivery mode. However it is not the only method in use: other delivery possibilities, such as broadcast (one-to-all), multicast (one-to-many) and anycast (one-to-one-of-many) are available. Partridge, Mendez, and Milliken (1993) proposed the host anycasting service for the first time in RFC 1546. The basic idea behind the anycast networking paradigm is to separate the service identifier from the physical host, and enable the service to act as a logical entity of the network. This idea of anycasting can be achieved in different layers (e.g., network and application layers) and they have both strengths and weaknesses as well. We focus on network-layer anycasting in this article, where a node sends a packet to an anycast address and the network will deliver the packet to at least one, and preferably only one of the competent hosts. This approach makes anycasting a kind of group communication in that a group of hosts are specified for a service represented by an anycast address and underlying routing algorithms are supposed to find out the appropriate destination for an anycast destined packet.

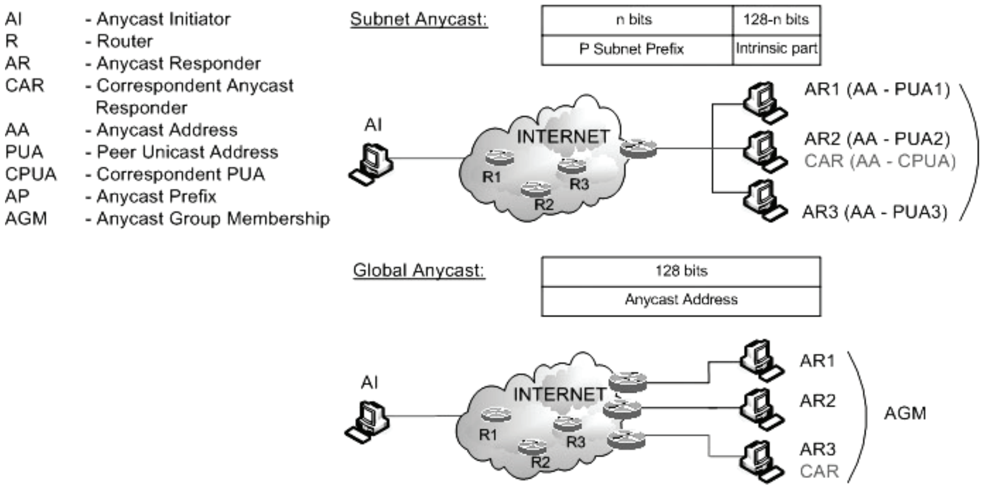
OVERVIEW OF IPV6 ANYCASTING

RFC 1546 introduced an experimental anycast address for IPv4, but in this case the anycast addresses were distinguishable from unicast addresses. IPv6 adopted the paradigm of anycasting as one of the basic and explicitly included services of IP and introduced the new anycast address besides the unicast and multicast addresses (Deering & Hinden, 1998). IPv6 anycast addresses were designed to allow reaching a single interface out of a group of interfaces. The destination node receiving the sent packets is the "nearest" node. The distance is dependent on the metric of the underlying routing protocol. In case of IPv6, an anycast address is defined as a unicast address assigned to more than one interface, so anycast addresses can not be distinguished from unicast addresses: they both share the same address space. Therefore the beginning part of any IPv6 anycast address is the network prefix. The longest *P* prefix identifies the topological region in which all interfaces are belonging to that anycast address reside. In the region identified by *P*, each member of the anycast membership must be handled as a separate entry of the routing system. Based on the length of *P*, IPv6 anycast can be categorized into two types: subnet anycast and global anycast. Hashimoto, Ata, Kitamura, and Murata (2005) summarized all that issues and defined the main terminology of IPv6 anycasting (Figure 1).

Hinden and Deering (2003) declared some restrictions concerning the further usage of the anycast addressing paradigm. The main purpose for setting these limitations was to keep the usage of anycast addresses under control until enough experience has been gathered in order to fit this new scheme to the existing structure of the Internet. These restrictions are now being eased that research could find appropriate solution for them (Abley, 2005). The biggest concern that had to be dealt with was routing since anycast packets (packets with an anycast address in the destination

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. IPv6 anycast terminology basics


AI - Anycast Initiator
R - Router
AR - Anycast Responder
CAR - Correspondent Anycast Responder
AA - Anycast Address
PUA - Peer Unicast Address
CPUA - Correspondent PUA
AP - Anycast Prefix
AGM - Anycast Group Membership
Subnet Anycast:
n bits
128-n bits
P Subnet Prefix
Intrinsic part
AI
INTERNET
R1
R2
R3
AR1 (AA - PUA1)
AR2 (AA - PUA2)
CAR (AA - CPUA)
AR3 (AA - PUA3)
AGM
Global Anycast:
128 bits
Anycast Address
AI
INTERNET
R1
R2
R3
AR1
AR2
AR3
CAR
AGM

field) might be forwarded to domains with different prefixes, as anycast receivers might be distributed all over the Internet. As a result a scalable and stable routing solution for anycasting is necessary.

Routing Protocols for IPv6 Anycasting

The current IPv6 standards do not define the anycast routing protocol, although the routing is one of the most important elements of network-layer anycasting. There is a quite small amount of literature about practical IPv6 anycasting. Park and Macker (1999) proposed and evaluated anycast extensions of link-state routing algorithm and distance-vector routing algorithm. Xuan, Jia,, Zhao, and Zhu (2000) proposed and compared several routing algorithms for anycast. Eunsoo Shim (2004) proposed an application load sensitive anycast routing method (ALSAR) and analyzed the existing routing algorithms in his PhD thesis. Doi, Ata, Kitamura, and Murata (2004) summarized the problems and possible solutions regarding the current specifications for IPv6 anycasting and proposed an anycast routing architecture based on seed nodes, gradual deployment and the similarities to multicasting. Based on their work, Matsunaga, Ata, Kitamura,and Murata (2005) designed and implemented three IPv6 anycast routing protocols (AOSPF—anycast open shortest path first, ARIP—anycast routing information protocol and PIA-SM—protocol independent anycast - sparse mode) based on existing multicast protocols.

The recent studies are focusing on subnet anycast routing protocols since they offer various possibilities for research while global anycast routing still faces scalability problems to be solved. The recently introduced anycast routing protocols all share a common ground as they are all based on multicast routing protocols because of the similarities of the two addressing schemes.

Unfortunately it does not fit the scope of this document to introduce each anycast routing protocol one-by-one although it is important to present the main idea that lies beneath all these protocols. The principal task to be performed is to discover all the anycast capable routers and nodes in the network: this can happen by flooding (as in case of AOSPF) or discovery methods (e.g., PIA-SM). The next, and maybe the most important step, is to maintain an up-to-date anycast routing table so all possible receivers could be reached in case of need. The easiest way to keep the routing entries up-to-date is to maintain a so-called Anycast Group Membership (Figure 1) where the anycast hosts can sign in or out when joining or leaving a certain anycast group designated by its anycast address.

APPLICATIONS OF ANYCASTING

Since the introduction of IPv6 anycast only a few applications have emerged using these addresses. It is mainly because the flexibility of the anycasting paradigm has not yet been widespread in the public. An excellent survey of the IPv6 anycast characteristics and applications was made by Weber and Cheng, 2004; Doi, Ata, Kitamura, and Murata, 2004; Matsunaga, Ata, Kitamura, and Murata (2005), where the authors describe many advantages and possible applications of anycasting. These applications can be classified into the following main types.

Main Application Schemas

The most popularly known application of anycast technology is helping the communicating nodes in selection of service providing servers. In the *server selection* approach the client host can choose one of many functionally identical servers.

The anycast server location and selection method could be a simple and transparent technique since the same address can be used from anywhere in the network, and the anycast routing would automatically choose the best destination for the client.

Anycast addresses can also be useful in discovering and locating services. In case of *service discovery,* the clients just need to know only one address: they can communicate with an optimal (e.g., minimum delay) host selected from the anycast group and easily discover the closest provider. This is especially beneficial in case of dynamically and frequently changing environments such as mobile ad-hoc systems. Services based on this characteristic can be acquired easily and optimally by the mobile clients through network-layer anycasting.

Application Scenarios

The most important advantage of network-layer anycasting is its ability to provide a simple mechanism where the anycast initiator (Figure 1) can receive a specific service without exact information about the server nodes and networks. Moreover the whole procedure is totally transparent: the clients do not need to know whether the server's address is unicast or anycast, because anycast addresses are syntactically indistinguishable from unicast addresses. Only servers have additional knowledge about their explicitly configured anycast addresses. The main application schemas and the application scenarios below are demonstrating the possibilities of the anycasting communication paradigm.

With the help of IPv6 anycasting *local information services* (e.g., emergency calls) can be given by getting each node to communicate with the appropriate server to the node's actual location. This kind of application is very useful in a mobile environment where nodes move from one network to another while resorting a given service.

By assigning a well-known anycast address to widespread applications, we can achieve *host auto-configuration.* The clients can use these services without knowing the appropriate unicast address of the server. The clients can utilize these applications everywhere only by specifying the service's well-known anycast address. For example, DNS resolvers no longer have to be configured with the unicast IP addresses for every host in every network if a standardized anycast address is built in the hardware or software, end users can get the service without configuration.

Improving the system reliability is another good example of IPv6 anycasting. Anycast communication grants multiple numbers of hosts with the same address and by increasing the number of hosts *load balancing, service redundancy* and *DoS attack avoidance* can be achieved based on the routing mechanism where anycast requests are fairly forwarded.

In a widely distributed environment (like a peer-to-peer architecture) services can construct a logical topology above the physical network. This logical topology can be based on anycast addresses. When a client wants to participate, it specifies the anycast address of the logical level in order to join in the logical network. In such a way, one of the participating nodes will become the *gate of the logical network* determined by the underlying anycast routing protocol.

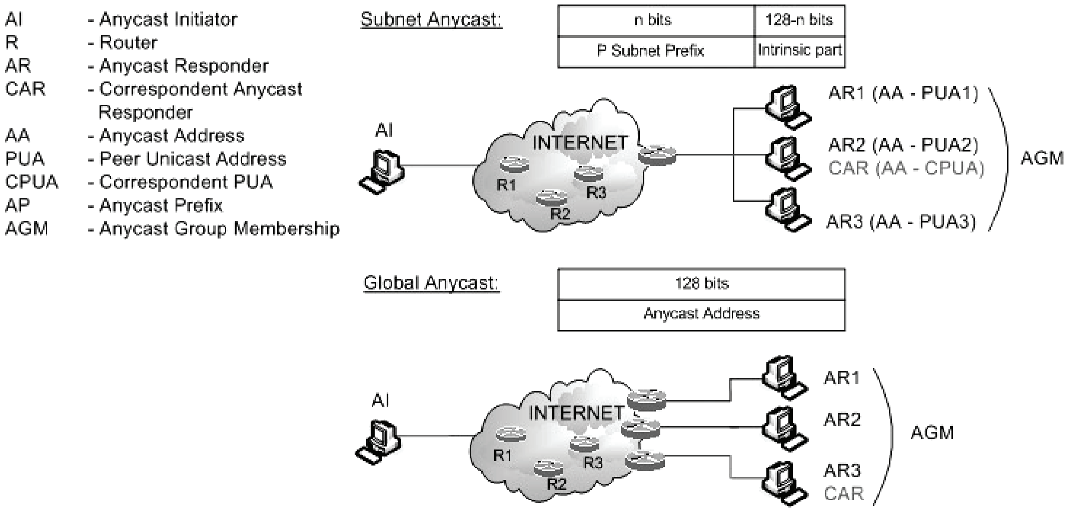
As we can see, there are some real promising application scenarios of IPv6 anycasting. However there is only one standardized anycast application these days, called *dynamic home agent address discovery*. In Mobile IPv6 the home agents (HA) have an anycast address, since the HA may change while the mobile terminal is not attached to its home network. Therefore a mobile node should use the anycast address of the home agents to reach one HA out of the set of home agents on its home link.

EMERGING APPLICATION: ANYCAST-BASED MICROMOBILITY

In Mobile IPv6 every mobile node (MN) is identified by its home address (HA), totally independently of where it is located in the network. When a MN is away from its home network (HN), it gets a new care-of-address (CoA). The IPv6 packets sent to the mobile node's HA will be routed to the mobile node's new CoA (Johnson, Perkins, & Arkko, 2004). Although Mobile IPv6 is capable of handling global mobility of users, it has shortcomings in supporting low latency and packet loss—required by real time multimedia services—during handover. To improve handover performance, the movement of a mobile node inside a subnet has to be dealt locally, by hiding intra-domain movements. As a result, the number of signaling messages reduced and the handover performance improved. Inside such a local subnet—called micromobility domain—the terminal receives a temporal IP address, which is valid throughout the subnet, and can be used a temporal CoA for the HA while the mobile terminal is located in the micromobility domain. Inside the micromobility domain, micromobility protocols are responsible for the proper routing of packets intended to the mobile hosts (Saha, Mukherjee, Misra, & Chakraborty, 2004). Leaving the micromobility domain, Mobile IPv6 provides global mobility management.

We have developed a new type of anycast application based on the main characteristics and the new research achievements of IPv6 anycasting in order to provide micromobility support in a standard IPv6 environment. In our proposed scheme, anycast addresses are used to identify mobile IPv6 hosts entering a micromobility domain while the underlying anycast routing protocol is used to maintain the anycast address routing information exchange. As a result the care-of-address obtained if the mobile terminal moves into a micromobility area is an anycast address. According to our proposal an anycast address identifies a single mobile

Figure 2. Entering a foreign micromobility domain

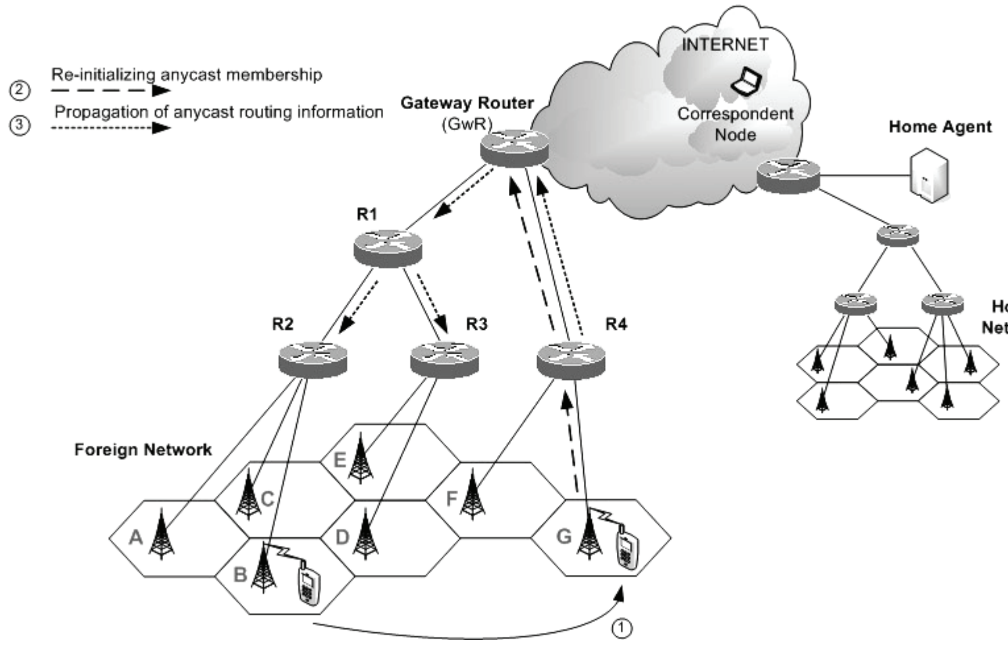

AI - Anycast Initiator
R - Router
AR - Anycast Responder
CAR - Correspondent Anycast Responder
AA - Anycast Address
PUA - Peer Unicast Address
CPUA - Correspondent PUA
AP - Anycast Prefix
AGM - Anycast Group Membership
Subnet Anycast:
n bits
128-n bits
P Subnet Prefix
Intrinsic part
AI
INTERNET
R1
R2
R3
AR1 (AA - PUA1)
AR2 (AA - PUA2)
CAR (AA - CPUA)
AR3 (AA - PUA3)
AGM
Global Anycast:
128 bits
Anycast Address
AI
INTERNET
R1
R2
R3
AR1
AR2
AR3
CAR
AGM

node. Therefore IP packets sent to the CoA of the mobile terminal have no chance to reach another "nearest" mobile node, since in this sense anycast addresses identifying mobile nodes are unique. The mobile node with a unique anycast care-of-address matches the correspondent anycast responder (CAR) in anycasting terminology. Also it has to be noted that in case of anycast address-based mobility there is no need for a peer unicast address since the CoA obtained is unique. The reason why unique anycast address is used instead of unicast address is the fact that anycast addresses are valid in the whole micromobility domain. Therefore the same anycast address can not be assigned to a second mobile node in a given micromobility domain.

The mobile node with a unique anycast address forms a virtual group. The members of this virtual group are the possible positions of the mobile node in the micromobility domain (that equals the validity area of the anycast address defined by the anycast *P* prefix) and the "nearest" mobile equipment is at the actual position of the mobile node. Therefore the mobile node remains reachable at any time (Figure 2). The purpose of using anycast address as an identifier for mobile nodes is that routing and handover management can be simplified with the help of changing the routing metrics. With the proper selection of the *P* prefix, the size of the virtual anycast group (VAG) can be adjusted easily. The virtual anycast group equals anycast group membership (AGM), while the virtual copies of the mobile node match the anycast responders. The operation of the anycast addressing-based mobility has to be investigated in case of different scenarios.

In the first scenario the mobile terminal leaves its current domain (e.g., its home network) and enters (1) another local administrative mobility domain (a new micromobility domain), as seen in Figure 2. In such case the mobile node first of all obtains (2)—with the help of IPv6 address autoconfiguration method—a unique anycast address that is valid in the whole area due to the properly set *P* prefix of the anycast address. As a result, the source address can be a unique anycast address since the source of a packet can be identified unequivocally. After getting the unique anycast care-of-address, the mobile node has to build the binding towards its home agent; therefore a binding procedure (3) is started by sending a binding update message. Next the mobile terminal has to initiate its membership in the virtual anycast group (VAG) of the new micromobility domain by having its anycast CoA (4). On receiving an anycast group membership report message the anycast access-router starts to propagate the new routing information by creating special routing information messages and sending it towards its adjacent routers. Based upon the underlying anycast routing protocol, each router in the new micromobility domain will get an entry in their routing table on how to reach the mobile terminal. Since each routing entry has a timeout period, thus the mobile node should send the membership report mes-

Figure 3. Moving in a given micromobility domain



sage periodically to maintain its routing entry. The updating time of the routing entry should be defined according to the refresh interval of the routing entries.

In the second scenario (Figure 3), the mobile node moves in a given micromobility area (1). At the new wireless point of attachment the mobile terminal has to notify the new access router about its new location. This updating process can be done, for example, with the help of data packets of an active communication. In this case the new access router notices that packets with the anycast address in the source address field are being sent over one of its interfaces (2) (the access router checks the direction where it receives the anycast-sourced packets). According to the anycast routing protocol the access router has an entry in its routing table regarding this source anycast address. Therefore the router modifies the entry regarding the anycast address of the mobile node so that the new entry forwards the packets towards their new destination (the interface from which it has received the packet with the anycast address in the source address field), the actual location of the mobile terminal. The access router also has to initiate anycast routing information exchange (3).

Our approach gives a unique viewpoint on applications of IPv6 anycasting: introduces a new solution for micromobility management based on the IPv6 anycast addresses. The proposed method fits to the Mobile IPv6 standard and works efficiently in micromobility environment, while reducing the volume of control messages during the mobile operation and resulting in more seamless handover. The procedure can be realized without new protocol stacks, because the method is based only on the built-in features of IPv6 standard. The anycast-based micromobility can work on any mobility-supporting IPv6 system.

FUTURE TRENDS

In this article we have tried to give you an overview of the main issues that are being tackled by the ongoing research. The focus of the recent research is to construct an anycast routing protocol that is capable of handling large amounts of anycast hosts with reasonably low overhead generated by the routing system. At the moment there are multiple candidate protocols that could fulfill all the requirements set for the anycast routing protocol, while the standardization of these protocols is on the way.

It is also important to take a look on the trends that can be found among the applications. One can easily see that various applications could benefit from the properties of anycast addressing, therefore more and more applications emerge for exploit these possibilities.

First of all, it should be highlighted that application of the anycast addressing scheme is closely related to introduction of IPv6 into today's network, therefore until the usage of IPv6 gets more widespread, the scope of anycasting is also limited. In accordance with the present trends the vision for the anycasting looks bright, since as soon as there will be standardized routing protocols more and more application will be able to use the advanced services of the anycast addressing. In our view micromobility management could be one of the driving applications using anycast addresses.

CONCLUSION

Our aim in this article was to present an overview of the usage of anycast addressing paradigm and also show a possible new usage of the anycast address introduced in IPv6. The proposed anycast-based micromobility scheme is fairly simple: the mobile node after joining a foreign network obtains a unique anycast care-of-address that is valid until the mobile terminal stays inside the micromobility area, no matter if the mobile node moves around. Our method uses the services of anycast routing protocols that are capable of routing the traffic towards the "nearest" node from the set of nodes having the same anycast address. Currently none of the existing anycast routing protocols have been widely adopted, due to the lack of standardization.

REFERENCES

Abley, J. (2005). *Anycast addressing in IPv6*. draft-jabley-v6-anycast-clarify-00.txt

Deering, S., & Hinden, R. (1998), *Internet Protocol Version 6 (IPv6)*. IETF RFC 2460.

Doi, S., Ata, S., Kitamura, H., & Murata M. (2004). IPv6 anycast for simple and effective service-oriented communications. *IEEE Communications Magazine*, 163-171.

Doi, S., Ata, S., Kitamura, H., & Murata, M. (2005). *Design, implementation and evaluation of routing protocols for IPv6 anycast communication*. In *19th International Conference on Advanced Information Networking and Applications AINA'05* (Vol. 2, pp. 833-838). Taiwan.

Eunsoo, S. (2004). *Mobility management in the wireless Internet*. PhD Thesis, Columbia University.

Hashimoto, M., Ata, S., Kitamura, H., & Murata, M. (2005). *IPv6 anycast terminolgy definition*. draft-doi-ipv6-anycast-func-term-03.txt

Hinden, R., & Deering, S. (2003). *.IP Version 6 Addressing Architecture*. IETF RFC 3513.

Johnson, D., Perkins, C., & Arkko, J. (2004). *Mobility support in IPv6*. IETF RFC 3775.

Macker, J. P., Park, V. D., & Corson, S. M. (2001). Mobile and wireless Internet services: Putting the pieces together. *IEEE Communications Magazine*, *39*(6), 148-155

Matsunaga, S., Ata, S., Kitamura, H., & Murata, M. (2005). *Applications of IPv6 Anycasting.* draft-ata-ipv6-anycast-app-01.txt.

Matsunaga, S., Ata, S., Kitamura, H., & Murata, M. (2005). Design and implementation of IPv6 anycast routing protocol: PIA-SM. In *19th International Conference on Advanced Information Networking and Applications (AINA'05)*, (Vol. 2, pp. 839-844). Taiwan.

Partridge, C., Mendez, T., & Milliken, W. (1993). *Host anycasting service*. IETF RFC 1546.

Saha, D., Mukherjee, A., Misra, I.S., & Chakraborty, M. (2004). Mobility support in IP: A survey of related protocols. *IEEE Network*, *18*(6), 34-40.

Park, V. D., & Macker, J .P. (1999). Anycast routing for mobile networking. In *MILCOM '99 Conference Proceedings*.

Weber, S., & Cheng, L. (2004). A survey of anycast in IPv6 Networks. *IEEE Communications Magazine*, 127-133

Xuan, D., Jia, W., Zhao, W., & Zhu, H. (2000). A routing protocol for anycast messages. *IEEE Transactions on Parallel and Distributed Systems*, *11*(6), 571-588

Applications Suitability on PvC Environments

Andres Flores
University of Comahue, Argentina

Macario Polo Usaola
Universidad de Castilla-La Mancha, Spain

INTRODUCTION

Pervasive computing (PvC) environments should support the *continuity* of users' daily tasks across dynamic changes of operative contexts. Pervasive or ubiquitous computing implies computation becoming part of the environment. Many different protocols and operating systems, as well as a variety of heterogeneous computing devices, are interrelated to allow accessing information anywhere, anytime in a secure manner (Weiser, 1991; Singh, Puradkar, & Lee, 2005; Ranganathan & Campbell, 2003).

According to the initial considerations by Weiser (1991), a PvC environment should provide the feeling of an enhanced natural human environment, which makes the computers themselves vanish into the background. Such a disappearance should be fundamentally a consequence not of technology but of human psychology, since whenever people learn something sufficiently well, they cease to be aware of it.

This means that the user's relationship to computation changes to an implicit human-computer interaction. Instead of thinking in terms of doing explicit tasks "*on the computer*"—creating documents, sending e-mail, and so on—on PvC environments individuals may behave as they normally do: moving around, using objects, seeing and talking to each other. The environment is in charge of facilitating these actions, and individuals may come to expect certain services which allow the feeling of "*continuity*" on their daily tasks (Wang & Garlan, 2000).

Users should be allowed to change their computational tasks between different operative contexts, and this could imply the use of many mobile devices that help moving around into the environment. As a result, the underlying resources to run the required applications may change from wide memory space, disk capacity, and computational power, to lower magnitudes. Such situations could make a required service or application inappropriate in the new context, with a likely necessity of supplying a proper adjustment. However, users should not perceive the surrounding environment as something that constraints their working/living activities. There should be a continuous provision of proper services or applications. Hence the environment must be provided with a mechanism for *dynamic applications suitability* (Flores & Polo, 2006).

PERVASIVE COMPUTING ENVIRONMENTS

In the field of PvC there is still a misuse of some related concepts, since often PvC is used interchangeably with ubiquitous computing and mobile computing. However, nowadays consistent definitions are identified in the literature as follows (Singh et al., 2005).

Mobile computing is about elevating computing services and making them available on mobile devices using the wireless infrastructure. It focuses on reducing the size of the devices so that they can be carried anywhere or by providing access to computing capacity through high-speed networks. However, there are some limitations. The computing model does not change considerably as we move, since the devices cannot seamlessly and flexibly obtain information about the context in which the computing takes place and adjust it accordingly. The only way to accommodate the needs and possibilities of changing environments is to have users manually control and configure the applications while they move—a task most users do not want to perform.

PvC deals with acquiring context knowledge from the environment and dynamically building computing models dependent on context. That is, providing dynamic, proactive, and context-aware services to the user. It is invisible to human users and yet provides useful computing services (Singh et al., 2005). Three main aspects must be properly understood (Banavar & Bernstein, 2002). First is the way people view mobile computing devices and use them within their environments to perform tasks. A device is a portal into an application/data space, not a repository of custom software managed by the user. Second is the way applications are created and deployed to enable such tasks to be performed. An application is a means by which a user performs a task, not a piece of software that is written to exploit a device's capabilities. And third is the environment and how it is enhanced by the emergence and ubiquity of new information and functionality. The computing environment is the user's information-enhanced physical surroundings, not a virtual space that exists to store and run software.

Ubiquitous computing uses the advances in mobile computing and PvC to present a *global computing environment* where seamless and invisible access to computing resources

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Vision of an enhanced physical environment by ubiquitous computing

is provided to the user. It aims to provide PvC environments to a human user as s/he moves from one location to another. Thus, it is created by sharing knowledge and information between PvC environments (Singh et al., 2005). Figure 1 shows the vision a user may have of a physical environment that is enhanced by ubiquitous computing.

Some approaches for PvC are concerned with interconnecting protocols from different hardware artifacts and devices, or solving problems of intermittent network connections and fluctuation on bandwidth. Therefore, their applications are quite general or low level, yet mainly related to communication tools which still requires a big effort for a user to accomplish a working task. Other approaches are focused on solving problems of prohibited access to information or even to a closed or restricted environment. If we consider that the environment is populated with an enormous amount of users, each intending accesses to different hardware and software resources, the security concerns increase proportionally (Kallio, Niemelä, & Latvakoski, 2004).

On the other side, there are approaches particularly concerned with providing higher level services more related to users tasks, in order to help them reduce the working effort (Roman, Ziebart, & Campbell, 2003; Becker & Schiele, 2003; Chakraborty, Joshi, Yesha, & Finin, 2006; Gaia Project, 2006; Aura Project, 2006). Most of them have been conceptualized with some sort of self-adjusted applications or by applications relying on basic services provided by the underlying platform (e.g., CORBA).

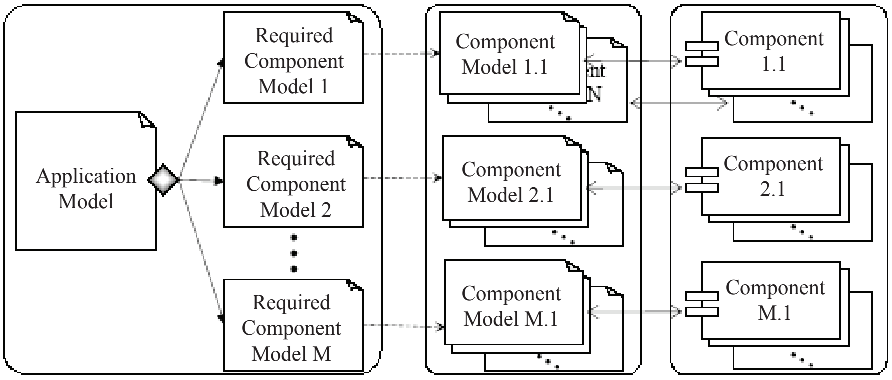
No matter how users need a transparent delivery of functionality, so they could have a sense of continued presence of the environment. Therefore, any unavailability of a required service implies that a user understand that the underlying environment cannot provide all that is needed, thus destroying the aspiration of transparency.

SUITABILITY FOR PERVASIVE APPLICATIONS

Functionality on a PvC environment is usually shaped as a set of aggregated components that are distributed among different computing devices. On changes of availability of a given device, the involved component behavior still needs to be accessible in the appropriate form according to the updated technical situation. This generally makes users be involved on a dependency with the underlying environment and increases the complexity of its internal mechanisms (Iribarne, Troya, & Vallecillo, 2003; Warboys et al., 2005).

Applications composed of dynamically replaceable components imply the need of an appropriate integration process according to component-based software development (CBSD) (Cechich, Piattini, & Vallecillo, 2003; Flores, Augusto, Polo, & Varea, 2004). For this, an application model may provide the specification of a required functionality in the form of the aggregation of component models, as can be seen in Figure 2. A component model provides a definition to

Figure 2. Connection of models and components to integrate an application



instantiate a component and its composition aspects through standard interactions and unambiguous interfaces (Cechich et al., 2003; Iribarne et al., 2003; Warboys et al., 2005). In order to assure the adequacy of a given component with respect to an application model, there is a need to evaluate its component model. Hence we present an assessment procedure which can be applied both on a development stage and also at runtime. The latter becomes necessary when the current technical situation makes unsuitable a given component demanding that a surrogate be provided.

The assessment procedure compares functional aspects from components against the specification provided by the application model, which is component oriented. Besides analyzing component services at a syntactic level, its behavior is also inspected, thus embracing semantic aspects. The latter is done by abstracting out the black box functionality hidden on components in the form of *assertions,* and also exposing its likely interactions by means of the *protocol of use,* which describes the expected order of use for its services (Flores & Polo, 2005)—also called choreography (Iribarne et al., 2003).

So far we have been experimenting with the addition of metadata for comparing behavioral aspects from components. Metadata has been used in several approaches as a technique to easy verification procedures (Cechich et al., 2003; Cechich & Polo, 2005; Orso et al., 2001). By adding meta-methods we may then retrieve detailed information concerning *assertions* and the *protocol of use,* which somehow implies a component adaptation, particularly referred to as the instrumentation mechanism (Flores & Polo, 2005).

The assessment procedure is described by means of a set of conditions which must be satisfied according to certain thresholds. Different techniques are applied to achieve the required evaluations. Compatibility on both assertions and the usage protocol is carried out by generating Abstract Syntax Trees and applying some updated algorithms, which were originally developed to detect similar pieces of code (*clones*) on existing programs. Such compatibility analysis is based on the following consideration. Post-conditions (for example) on services from two similar components necessarily should relate to a similar structure and semantic. Hence, they could be thought of as one being a clone of the other (Flores & Polo, 2005).

All such techniques applied on our assessment procedure allow the accomplishment of a consistent mechanism to assure a fair component integration. As PvC environments imply many challenges, the whole integration process is based on considering all aspects concerning reliability.

We may make use of a simple example to illustrate the way a functionality is composed from distributed disparate components and understand how the assessment procedure may help to assure the suitability of a certain involved component.

Suppose we represent a PvC environment for a museum, which includes a tour guide application for proposing different paths according to the user's dynamic choices. When the user enters the museum, s/he may carry a computing device (a PDA or a smart phone), and an automatic detection is done in order to identify and connect the device to the environment. As the user walks by each art piece (e.g., painting or sculpture), descriptions and information of particular interest to the user are displayed on the PDA or spoken through the phone. Figure 3 shows a likely scenario of the presented case study.

A related application could allow creating an album with images of some art pieces the user has visited. To obtain the images the user will probably have to pay a fee, for example, when s/he intends to leave the museum. The album organizer application—maybe downloaded into a user's notebook recognized by the environment—may allow creating a sort of document with images and some notes written by the user. Notes could be stored on separated text files and bind to the document by means of hypertext links. Thus every time the user needs to write or edit a note, a proper editor is provided.

Figure 3. PvC environment for a museum

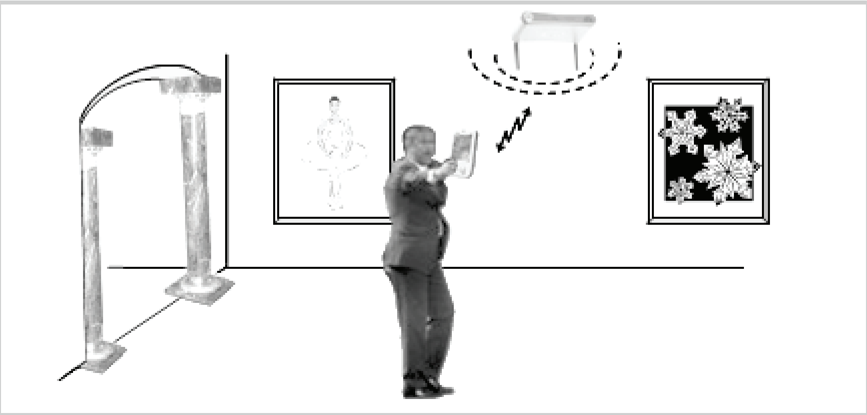
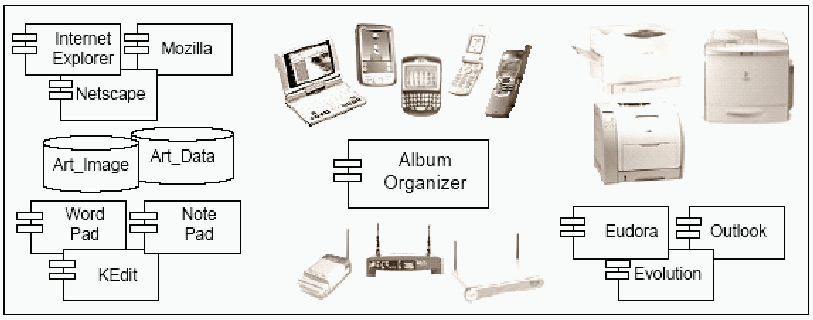


Figure 4. Distributed components for the album organizer application



The user may also be allowed to print a selection of pages of the document, or even send the created album by e-mail in case s/he does not have a device to carry those files.

If we focus on the album organizer application, we might analyze the potential required components. There could be an album organizer component to represent the main logic of the application. This component could have an ad-hoc sophisticated album visual editor or it could be a Web-style editor in which is additionally required a generic Web browser. The visual editor also depends on the actual used device. In order to make notes, different components could be used as a simple sort of Notepad, or other replacements like WordPad or similar applications according to the underlying software platform. For sending e-mail, applications like Outlook or Eudora could be used, and to provide a printer service, different kinds of printers and ad hoc wireless sensors should be available. Another component is concerned with the database for images and descriptions of art pieces. Figure 4 shows a diagram with the likely comprised components and devices for the album organizer application.

Suppose a user needs to write a note by using a notebook which runs a Linux platform. One available text editor is KEdit. The environment then evaluates this component so to ensure it is appropriate to fulfill the task. Following can be seen the interfaces of both the KEdit component and the required component model named TextEditor. In order to assure the compatibility with the surrogate, the assessment procedure is applied to analyze if the degree of similarity raises a proper level.

Since implementation alternatives are fairly important, we have properly explored some of them. For example both Microsoft .NET and Enterprise Java Beans allow the addition of metadata to components. Information about components—their structure and the state of their corresponding instances—can be retrieved at runtime by using *reflection* on both environments as well. Particularly .NET includes the possibility of adding *attributes,* which is a special class intended to provide additional information about some design element as a class, a module, a method, a field, and so on (Flores & Polo, 2005).

Hence, we have implemented on the .Net technology the current state of our approach (Flores & Polo, 2006). Though simple, this prototype gives us rewarding data on possibilities to make our proposals concrete. All the applied techniques are selected according to our goal of achieving consistent mechanisms to assure a fair component integration. As we proceed with our work, reliability is mainly considered, since we focus the whole integration process for those challenging systems as PvC environments.

FUTURE TRENDS

Our work continues by completing the coverage of functional aspects for components, describing the components replacement mechanism and then focusing on the non-functional aspects, particularly on quality of service. For this we are analyzing the use and extending the schemas from Iribarne et al. (2003) which provide a consistent format for specifications of components by means of XML. The approach covers all of the aspects from components: functional, non-functional, and commercial.

As some authors have pointed out (Chen, Finin, & Joshi, 2005; Ranganathan & Campbell, 2003), temporal aspects could also be helpful to achieve a more accurate component integration process. This may give the chance to analyze whether a component can fit the requirements when time conditions are also included in the set of updated requirements upon a change on the user's context of operation.

Selection of appropriate methods, techniques, and languages must be accurately accomplished upon the concern of a reliable mechanism. This is the emphasis of our next development in this area.

REFERENCES

Aura Project. (2006). *Aura Project Web site.* Retrieved from http://www-2.cs.cmu.edu/ aura/

Banavar, G., & Bernstein, A. (2002). Issues and challenges in ubiquitous computing: Software infrastructure and design challenges for ubiquitous computing applications. *Communications of the ACM, 45*(12).

Becker, C., & Schiele, G. (2003). Middleware and application adaptation requirements and their support in pervasive computing. *Proceedings of IEEE ICDCSW* (pp. 98-103), Providence, RI.

Cechich, A., & Polo, M. (2005). COTS component testing through aspect-based metadata. In S. Beydeda & V. Gruhn (Eds.), *Building quality into components—testing and debugging.* Berlin: Springer-Verlag.

Cechich, A., Piattini, M., & Vallecillo, A. (Eds.). (2003). *Component-based software quality: Methods and techniques* (LNCS 2693). Berlin: Springer-Verlag.

Chakraborty, D., Joshi, A., Yesha, Y., & Finin, T. (2006). Toward distributed service discovery pervasive computing environments. *IEEE Transactions on Mobile Computing, 5*(2).

Chen, H., Finin, T., & Joshi, A. (2004). Semantic Web in the context broker architecture. *Proceedings of IEEE PerCom*, Orlando, FL, (pp. 277-286).

Flores, A., & Polo, M. (2005, June 27-30). Dynamic component assessment on PvC environments. *Proceedings of IEEE ISCC*, Cartagena, Spain, (pp. 955-960).

Flores, A., & Polo, M. (2006, May 23). An approach for applications suitability on pervasive environments. *Proceedings of IWUC* (held at ICEIS). Paphos, Cyprus: INSTICC Press.

Flores, A., Augusto, J. C., Polo, M., & Varea, M. (2004, October 10-13). Towards context-aware testing for semantic interoperability on PvC environments. *Proceedings of IEEE SMC*, The Hague, The Netherlands, (pp. 1136-1141).

Gaia Project. (2006). *Gaia Project Web site.* Retrieved from http://www.w3.org/2001/sw

Iribarne, L., Troya, J., & Vallecillo, A. (2003). A trading service for COTS components. *The Computer Journal, 47*(3).

Kallio, P., Niemelä, E., & Latvakoski, J. (2004). UbiSoft—pervasive software. *Research Notes, 2238.* Finland: VTT Electronics. Retrieved from www.vtt.fi/inf/pdf/tiedotteet/2004/T2238.pdf

Orso, A., Harrold, M.J., Rosenblum, D., Rothermel, G., Do, H., & Sofía, M.L. (2001). Using component metacontent to support the regression testing of component-based software. *Proceedings of IEEE ICSM*, Florence, Italy, (pp. 716-725).

Ranganathan, A., & Campbell, R. (2003). An infrastructure for context-awareness based on first order logic. *Personal and Ubiquitous Computing, 7,* 353-364.

Roman, M., Ziebart, B., & Campbell, R. (2003). Dynamic application composition: Customizing the behavior of an active space. *Proceedings of IEEE PerCom.*

Singh, S., Puradkar, S., & Lee, Y. (2005). Ubiquitous computing: Connecting pervasive computing through semantic Web. *Journal of ISeB.* Berlin: Springer-Verlag.

Wang, Z., & Garlan, D. (2000). *Task-driven computing.* Technical Report No. CMU-CS-00-154, School of Computer Science, Carnegie Mellon University, USA. Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/2000/abstracts/00-154.html

Warboys, B., Snowdon, B., Greenwood, R. M., Seet, W., Robertson, I., Morrison, R., Balasubramaniam, D., Kirby, G., & Mickan, K. (2005). An active-architecture approach to COTS integration. *IEEE Software, 22*(4), 20-27.

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 94-104. Retrieved from http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html

KEY TERMS

Component-Based Software Development (CBSD): A development paradigm where software systems are developed from the assembly or integration of software components.

Component Model: A specification that describes how to instantiate or build a software component, and gives guidelines for its binding to other software components by means of standard interactions or communication patterns and unambiguous interfaces.

Mobile Computing: Small wireless devices that can be carried anywhere, allowing a computing capacity through wireless networks. Also known as nomadic computing.

Pervasive Computing (PvC): Enhancement of the physical surroundings by providing and adapting mobile computing according to the user's needs. Also known as ambient intelligent.

Software Component: A unit of independent deployment that is ready "off-the-shelf" (OTS), from a commercial source (COTS) or reused from another system (in-house or legacy systems). It is usually self-contained, enclosing a collection of cooperating and tightly cohesive objects, thus providing a significant aggregate of functionality. It is used "as it is found" rather than being modified, may possibly execute independently, and can be integrated with other components to achieve a required bigger system functionality.

Ubiquitous Computing: Provides PvC environments to a human user as s/he moves from one location to another. It allows sharing knowledge and information between PvC environments. Also known as Global Computing.

Wireless Device: A small computer that is reduced in size and in computing power, can be carried everywhere, and provides voice, data, games, and video applications. The most familiar is the mobile or cell phone, then Palm Pilot and its handheld descendent, the PDA (personal digital assistant), a great evolution because of its large amount of new applications. Laptop or notebook and tablet PCs are also well-known wireless devices.

Wireless Network: Offers mobility and elimination of unsightly cables, by the use of radio waves and/or microwaves to maintain communication channels between computers. It is an alternative to wired networking, which relies on copper and/or fiber optic cabling between network devices. Popular wireless local area networking (WLAN) products conform to the 802.11 "Wi-Fi" standards.

A Bio-Inspired Approach for the Next Generation of Cellular Systems

Mostafa El-Said
Grand Valley State University, USA

INTRODUCTION

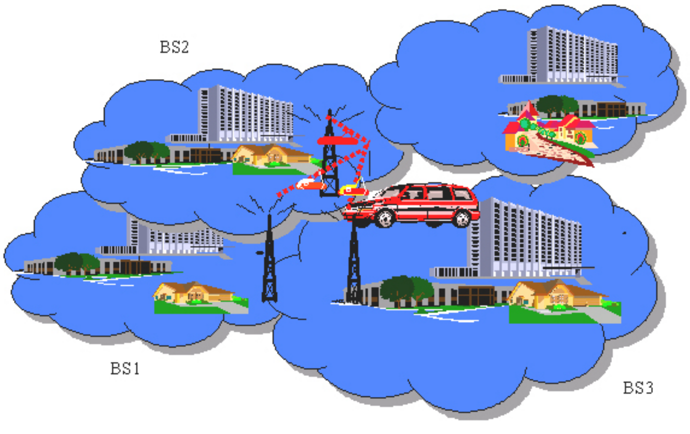
In the current 3G systems and the upcoming 4G wireless systems, *missing neighbor pilot* refers to the condition of receiving a high-level pilot signal from a Base Station (BS) that is not listed in the mobile receiver's neighbor list (LCC International, 2004; Agilent Technologies, 2005). This pilot signal interferes with the existing ongoing call, causing the call to be possibly dropped and increasing the handoff call dropping probability. Figure 1 describes the missing pilot scenario where BS1 provides the highest pilot signal compared to BS1 and BS2's signals. Unfortunately, this pilot is not listed in the mobile user's active list.

The horizontal and vertical handoff algorithms are based on continuous measurements made by the user equipment (UE) on the Primary Scrambling Code of the Common Pilot Channel (CPICH). In *3G systems,* UE attempts to measure the quality of all received CPICH pilots using the Ec/Io and picks a dominant one from a cellular system (Chiung & Wu, 2001; El-Said, Kumar, & Elmaghraby, 2003). The UE interacts with any of the available radio access networks based on its memorization to the neighboring BSs. As the UE moves throughout the network, the serving BS must constantly update it with neighbor lists, which tell the UE which CPICH pilots it should be measuring for handoff purposes. In *4G systems*, CPICH pilots would be generated from any wireless system including the 3G systems (Bhashyam, Sayeed, & Aazhang, 2000). Due to the complex heterogeneity of the 4G radio access network environment, the UE is expected to suffer from various carrier interoperability problems. Among these problems, the missing neighbor pilot is considered to be the most dangerous one that faces the 4G industry.

The wireless industry responded to this problem by using an inefficient traditional solution relying on using antenna downtilt such as given in Figure 2. This solution requires shifting the antenna's radiation pattern using a mechanical adjustment, which is very expensive for the cellular carrier. In addition, this solution is permanent and is not adaptive to the cellular network status (Agilent Technologies, 2005; Metawave, 2005).

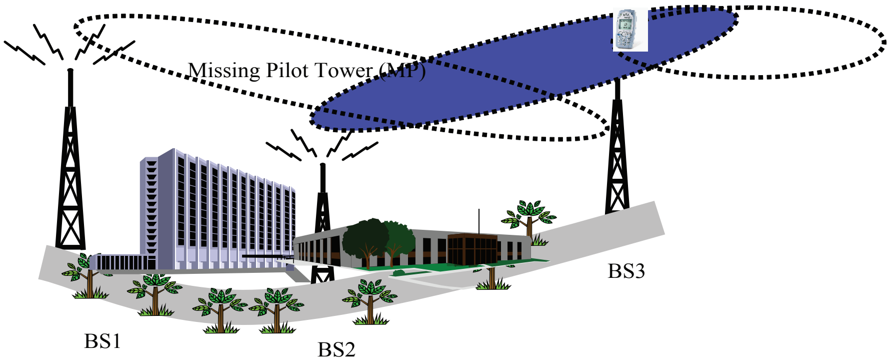
Therefore, a self-managing solution approach is necessary to solve this critical problem. Whisnant, Kalbarczyk, and Iyer (2003) introduced a system model for dynamically reconfiguring application software. Their model relies on considering the application's static structure and run-time behaviors to construct a workable version of reconfiguration software application. Self-managing applications are hard to test and validate because they increase systems complexity (Clancy, 2002). The ability to reconfigure a software application requires the ability to deploy a dynamically hardware infrastructure in systems in general and in cellular systems in particular (Jann, Browning, & Burugula, 2003).

Figure 1. Missing pilot scenario



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Missing pilot solution: Antenna downtilt



Konstantinou, Florissi, and Yemini (2002) presented an architecture called NESTOR to replace the current network management systems with another automated and software-controlled approach. The proposed system is inherently a rule-based management system that controls change propagation across model objects. Vincent and May (2005) presented a decentralized service discovery approach in mobile ad hoc networks. The proposed mechanism relies on distributing information about available services to the network neighborhood nodes using the analogy of an electrostatic field. Service requests are issued by any neighbor node and routed to the neighbor with the highest potential.

The autonomic computing system is a concept focused on adaptation to different situations caused by multiple systems or devices. The IBM Corporation recently initiated a public trail of its Autonomic Toolkit, which consists of multiple tools that can be used to create the framework of an autonomic management system. In this article, an autonomic engine system setting at the cellular base station nodes is developed to detect the missing neighbor (Ganek & Corbi, 2003; Haas, Droz, & Stiller, 2003; Melcher & Mitchell, 2004). The autonomic engine receives continuous feedback and performs adjustments to the cell system's neighboring set by requiring the UE to provide signal measurements to the serving BS tower (Long, 2001).

In this article, I decided to use this toolkit to build an autonomic rule-based solution to detect the existence of any missing pilot. The major advantage of using the IBM autonomic toolkit is providing a common system infrastructure for processing and classifying the RF data from multiple sources regardless of its original sources. This is a significant step towards creating a transparent autonomic high-speed physical layer in 4G systems.

PROPOSED SOLUTION

The proposed AMS relies on designing an autonomic high-speed physical layer in the smart UE and the BS node. *At the UE side,* continuous CPICH pilot measurements will be recorded and forwarded to the serving BS node via its radio interface. *At the BS node,* a scalable self-managing autonomic engine is developed using IBM's autonomic computing toolkit to facilitate the mobile handset's vertical/horizontal handover such as shown in Figure 3. The proposed engine is cable of interfacing the UE handset with different wireless technologies and detects the missing pilot if it is existed.

The autonomic engine relies on a generic log adapter (GLA), which is used to handle any raw measurements log file data and covert it into a standard format that can be understood by the autonomic manager. Without GLA, separate log adapters would have be coded for any system that the autonomic manager interfaced with. The BS node will then lump all of the raw data logs together and forward them to the Generic Log Adapter for data classification and restructuring to the common base event format. Once the GLA has parsed a record in real time to common base event format, the autonomic manager will see the record and process it and take any action necessary by notifying the BS node to make adjustments to avoid the missing pilot and enhance the UE devices' quality of service.

PERFORMANCE MEASUREMENTS AND KEY FINDINGS

To test the applicability of the proposed solution, we decided to use the system's response time, AS's service rate for callers experiencing missing pilot problem, and the performance

Figure 3. Autonomic base station architecture

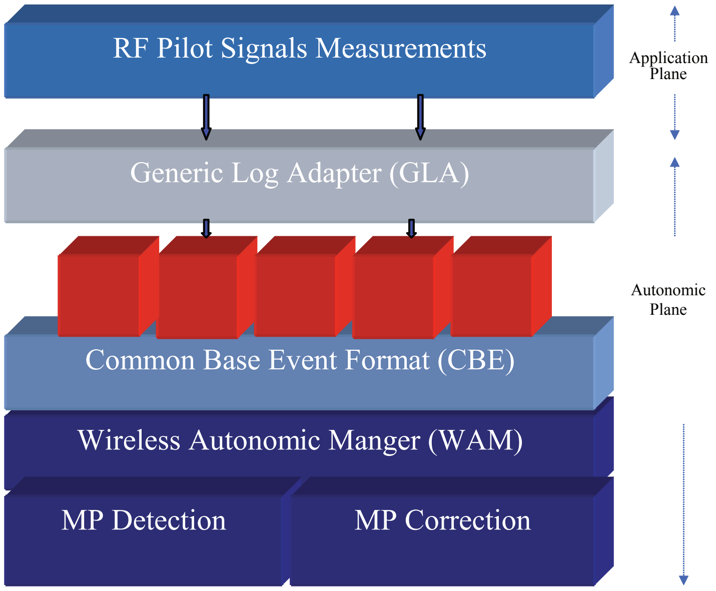

RF Pilot Signals Measurements
Application Plane
Generic Log Adapter (GLA)
Autonomic Plane
Common Base Event Format (CBE)
Wireless Autonomic Manger (WAM)
MP Detection
MP Correction

Table 1. Summary of the system performance analysis

	Log File Size in (# Records)	System Response Time in (Sec)	Processing Rate by the Base Station in (Records/Sec)
Trial Experiment 1	985	145	6.793103448
Trial Experiment 2	338	95	3.557894737
Trial Experiment 3	281	67	4.194029851
Trial Experiment 4	149	33	4.515151515
Average Processing Rate by the Base Station in (Records/Sec)	**4.765044888**		
Base Station Service Rate For callers experiencing missing pilot problem (Records/Sec)	**5.3**		
Performance Gain	**1.112266542**		

gain as performance metrics. Also, we developed a Java class to simulate the output of a UE in a heterogeneous RF access network. Table 1 summarizes the simulation results for four simulation experiments with different log files size.

The results shown in Table 1 comply with the design requirements for the current 3G system. This is illustrated in the following simple example.

DESIGN REQUIREMENTS FOR 3G SYSTEMS

- The 3G cell tower's coverage area is divided into three sectors, with each sector having (8 traffic channel * 40 call/channel = 320 voice traffic per sector) and (2 control channels * 40 callers/channels = 80 control traffic per sector).
- The overlapped area between towers (handoff zone) occupies 1/3 of the sector size and serves (1/3 of 320

= 106 callers (new callers and/or exciting ones)). If we consider having the UE report its status to the tower every 5 seconds, we could potentially generate 21.2 records in 1 second.

- It is practical to assume that 25% of the 21.2 reports/second accounts for those callers that may suffer from the missing pilot problem—that is, the tower's service rate for missing neighbor pilot callers is 21.2/4 = 5.3 records/second. This is the threshold level used by the tower to accommodate those callers suffering from the missing pilot problem.

ANALYSIS OF THE RESULTS

- Response time is the time taken by the BS to process, parse the incoming log file and detect the missing neighbor pilot. It is equal to (145, 95, 67, and 33) for the four experiment trials. All values are in seconds.
- Processing rate by the base station is defined as the total number of incoming records divided by the response time in (records/second). It is equal to (6.7, 3.5, 4.1, and 4.5) for the four experiment trials.
- The UE reports a missing pilot problem with an average rate of 4.7 records/second.
- The base station's service rate for callers experiencing the missing pilot problem = 5.3 records/second.
- The performance gain is defined as:

$$\frac{\text{Base Station Service Rate For callers experiencing missing pilot problem in (Records/Sec)}}{\text{Average Processing Rate by the Base Station in (Records/Sec)}}$$

$= 5.3/4.7=1.1$

- Here it is obvious that the service rate (5.3 records/second) is greater than the UE's reporting rate to the base station node (4.7 records/second). Therefore, the above results prove that the proposed solution does not overload the processing capabilities of the BS nodes and can be scaled up to handle a large volume of data.

FUTURE TRENDS

An effective solution for the interoperability issues in 4G wireless systems must rely on an adaptive and self-managing network infrastructure. Therefore, the proposed approach in this article can be scaled to maintain continuous user connectivity, better quality of service, improved robustness, and higher cost-effectiveness for network deployment.

CONCLUSION

In this article, we have developed an autonomic engine system setting at the cellular base station (BS) nodes to detect the missing neighbor. The autonomic engine receives continuous feedback and performs adjustments to the cell system's neighboring set by requiring the user equipment (UE) to provide signal measurements to the serving BS tower. The obtained results show that the proposed solution is able to detect the missing pilot problem in any heterogeneous RF environment.

REFERENCES

Agilent Technologies. (2005). Retrieved October 2, 2005, from http://we.home.agilent.com

Bhashyam, S., Sayeed, A., & Aazhang, B. (2000). Time-selective signaling and reception for communication over multipath fading channels. *IEEE Transaction on Communications, 48*(1), 83-94.

Chiung, J., & Wu, S. (2001). Intelligent handoff for mobile wireless Internet. *Journal of Mobile Networks and Applications*, *6,* 67-79.

Clancy, D. (2002). *NASA challenges in autonomic computing. Almaden Institute 2002, IBM Almaden Research Center,* San Jose, CA.

El-Said, M., Kumar, A., & Elmaghraby, A. (2003). Pilot pollution interference cancellation in CDMA systems. *Special Issue of Wiley Journal: Wireless Communication and Mobile Computing on Ultra Broadband Wireless Communications for the Future, 3*(6), 743-757.

Ganek, A., & Corbi, T. (2003). The dawning of the autonomic computing era. *IBM Systems Journal, l42*(1), 5-19.

Haas, R., Droz, P., & Stiller, B. (2003). Autonomic service deployment in networks. *IBM Systems Journal, 42*(1), 150-164.

Jann, L., Browning, A., & Burugula, R. (2003). Dynamic reconfiguration: Basic building blocks for autonomic computing on IBM pSeries servers. *IBM Systems Journal, 42*(1), 29-37.

Konstantinou, A., Florissi, D., & Yemini, Y. (2002). Towards self-configuring networks. *Proceedings of the DARPA Active Networks Conference and Exposition* (pp. 143-156).

LCC International. (2004). Retrieved December 10, 2004, from http://www.hitech-news.com/30112001-MoeLLC.htm

Lenders, V., May, M., & Plattner, B. (2005). Service discovery in mobile ad hoc networks: A field theoretic approach. *Special Issue of Pervasive and Mobile Computing, 1,* 343-370.

Long, C. (2001). *IP network design.* New York: McGraw-Hill Osborne Media.

Melcher, B., & Mitchell, B. (2004). Towards an autonomic framework: Self-configuring network services and developing autonomic applications. *Intel Technology Journal, 8*(4), 279-290.

Metawave. (2005). Retrieved November 10, 2005, from http://www.metawave.com

Whisnant, Z., Kalbarczyk, T., & Iyer, R. (2003). A system model for dynamically reconfigurable software. *IBM Systems Journal, 42*(1), 45-59.

KEY TERMS

Adaptive Algorithm: Can "learn" and change its behavior by comparing the results of its actions with the goals that it is designed to achieve.

Autonomic Computing: An approach to self-managed computing systems with a minimum of human interference. The term derives from the body's autonomic nervous system, which controls key functions without conscious awareness or involvement.

Candidate Set: Depicts those base stations that are in transition into or out of the active set, depending on their power level compared to the threshold level.

Missing Neighbor Pilot: The condition of receiving a high-level pilot signal from a base station (BS) that is not listed in the mobile receiver's neighbor list.

Neighbor Set: Represents the nearby serving base stations to a mobile receiver. The mobile receiver downloads an updated neighbor list from the current serving base station. Each base station or base station sector has a unique neighbor list.

Policy-Based Management: A method of managing system behavior or resources by setting "policies" (often in the form of "if-then" rules) that the system interprets.

Virtual Active Set: Includes those base stations (BSs) that are engaged in a live communication link with the mobile user; they generally do not exceed three base stations at a time.

Brain Computer Interfacing

Diego Liberati
Italian National Research Council, Italy

INTRODUCTION

In the near future, mobile computing will benefit from more direct interfacing between a computer and its human operator, aiming at easing the control while keeping the human more free for other tasks related to displacement.

Among the technologies enabling such improvement, a special place will be held by brain computer interfacing (BCI), recently listed among the 10 emerging technologies that will change the world by the *MIT Technology Review* on January 19, 2004.

The intention to perform a task may be in fact directly detected from analyzing brain waves: an example of such capability has been for instance already shown trough artificial neural networks in Babiloni et al. (2000), thus allowing the switch of a bit of information in order to start building the control of a direct interaction with the computer.

BACKGROUND

Our interaction with the world is mediated through sensory-motor systems, allowing us both to acquire information from our surroundings and manipulate what is useful at our reach. Human-computer interaction ergonomically takes into account the psycho-physiological properties of such interaction to make our interactions with computers increasingly easy. Computers are in fact nowadays smaller and smaller without significant loss of power needed for everyday use, like writing an article like this one on a train going to a meeting, while checking e-mail and talking (via voice) to colleagues and friends.

Now, the center of processing outside information and producing intention to act consequently is well known to be our brain. The capability to directly wire neurons on electronic circuits is not (yet?) within our reach, while interesting experiments of compatibility and communication capabilities are indeed promising at least in vitro. At the other extreme, it is not hard to measure non-invasively the electromagnetic field produced by brain function by positioning small electrodes over the skull, as in the standard clinical procedure of electroencephalography.

Obviously, taking from outside a far-field outside measure is quite different than directly measuring the firing of every single motor neuron of interest: a sort of summing of all the brain activity will be captured at different percentages. Nonetheless, it is well known that among such a messy amount of signal, when repeating a task it is not hard to enhance the very signal related to task, while reducing—via synchronized averaging—the overwhelming contribution of all the other neurons not related to the task of interest. On this premise, Deecke, Grozinger, and Kornhuber (1976) have been able to study the so-called event-related brain potential, naming the onset of a neural activation preceding the task, in addition to the neural responses to the task itself.

Statistical pattern recognition and classification has been shown to improve such event-related detection by Gevins, Morgan, Bressler, Doyle, and Cutillo (1986).

A method to detect such preparatory potential on a single event basis (and then not needing to average hundreds of repetitions, as said before) was developed and applied some 20 years ago by Cerutti, Chiarenza, Liberati, Mascellani, and Pavesi (1988). One extension of the same parametric identification approach is that developed by del Millan et al. (2001) at the European Union Joint Research Center of Ispra, Italy, while a Bayesian inference approach has been complementary proposed by Roberts and Penny (2000).

BRAIN COMPUTER INTERFACING

Autoregressive with exogenous input parametric identification (Cerutti et al., 1988) is able to increase by some 20 dB the worst signal-to-noise ratio of the event-related potential with respect to the overwhelming background brain activity. Moreover, it provides a reduced set of parameters that can be used as features to perform post-processing, should it be needed.

A more sophisticated, though more computing-demanding, time variant approach based on an optimal so-called Kalman filer has been developed by Liberati, Bertolini, and Colombo (1991a). The joint performance of more than one task has also been shown to evoke more specific brain potential (Liberati, Bedarida, Brandazza, & Cerutti, 1991b).

Multivariable joint analysis of covariance (Gevins et al., 1989), as well as of total and partial coherence among brain field recordings at different locations (Liberati, Cursi, Locatelli, Comi, & Cerutti, 1997), has also improved the capability of discriminating the single potential related to a particular task (Liberati, 1991a).

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Artificial neural networks (Liberati, 1991b; Pfurtscheller, Flotzinger, Mohl, & Peltoranta, 1992; Babiloni et al., 2000) offered the first approach to the problem of so-called artificial intelligence, whose other methods of either soft computing, like the fuzzy sets made popular by Lofthi Zadeh, and the rule extraction like the one proposed by Muselli and Liberati (2002), are keen to be important post-processing tools for extracting real commands from the identified parameters.

In particular, the rule extraction approach proposed by Muselli and Liberati (2000) has the nice properties of processing the huge amount of data in a very fast quadratic time (and even in terms of binary operations), yielding both pruning of the redundant variables for discrimination (like not necessary recording points or time windows) and an understandable set of rules relating the residual variable of interest: this is thus quite useful in learning the BCI approach.

When the space of the feature is then confined to the few really salient, the Piece-Wise Affine identification recently developed (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003) and also applied in a similar context for instance to hormone pulse or sleep apnea detection (Ferrari-Trecate & Liberati, 2002), is keen to be a good tool to help refine the detection of such mental decision on the basis of the multivariate parametric identification of a multiple set of dynamic biometric signals. Here the idea is to cluster recorded data or features obtained by the described preprocessing in such a way to identify automatically both approximate linear relations among them in each region of interest and the boundaries of such regions themselves, thus allowing quite precise identification of the time of the searched switching.

FUTURE TRENDS

Integration of more easily recorded signals is even more promising for automatic processing of the intention of interacting with the computer, both in a context of more assisted performance in everyday life, as well as in helping to vicariate lost functions because of handicaps.

CONCLUSION

The task is challenging, though at a first glance it would even appear not so complex: it wants to discriminate at least a bit of information (like opening and closing a switch), and then sequentially, it would be possible to compose a word of any length.

The point is that every single bit of intention should be identified with the highest accuracy, in order to avoid too many redundancies, demanding time while offering safety.

REFERENCES

Babiloni, F., Carducci, F., Cerutti, S., Liberati, D., Rossini, P., Urbano, A., & Babiloni, C. (2000). Comparison between human and ANN detection of Laplacian-derived electroencephalographic activity related to unilateral voluntary movements. *Comput Biomed Res, 33,* 59-74.

Cerutti, S., Chiarenza, G., Liberati, D., Mascellani, P., & Pavesi, G. (1988). A parametric method of identification of the single trial event-related potentials in the brain. *IEEE Transactions of Biomedical Engineering, 35*(9), 701.

Deecke, L., Grözinger, B., & Kornhuber, H. (1976). Voluntary finger movements in man: Cerebral potentials and theory. *Biological Cybernetics, 23,* 99-119.

del Millan et al. (2001). Brain computer interfacing. In D. Liberati (Ed.), *Biosys: Information and control technology in health and medical systems.* Milan: ANIPLA.

Ferrari-Trecate, G., & Liberati, D. (2002). Representing logic and dynamics: The role of piecewise affine models in the biomedical field. *Proceedings of the EMSTB Math Modeling and Computing in Biology and Medicine Conference,* Milan.

Ferrari–Trecate, G., Muselli, M., Liberati, D., & Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica, 39,* 205-217.

Gevins, A., Morgan, N., Bressler, S., Doyle, J., & Cutillo, B. (1986). Improved event-related potential estimation using statistical pattern classification. *Electroenceph. Clin. Neurophysiol, 64,* 177.

Gevins, A., Bressler, S. L., Morgan, N. H., Cutillo, B., White, R. M., Greer, D. S., & Illes, J. (1989). Event-related covariances during a bimanual visuomotor task: Methods and analysis of stimulus and response-locked data. *Electroenceph. Clin. Neurophysiol, 74,* 58.

Liberati, D. (1991a), Total and partial coherence analysis of evoked brain potentials. *Proceedings of the 4th International Symposium on Biomedical Engineering,* Peniscola, Spain, (pp. 101-102).

Liberati, D. (1991b). A neural network for single sweep brain evoked potential detection and recognition. *Proceedings of the 4th International Symposium on Biomedical Engineering,* Peniscola, Spain, (pp. 143-144).

Liberati, D., Bertolini, L., & Colombo, D. C. (1991a). Parametric method for the detection of inter and intra-sweep variability in VEP's processing. *Med Biol Eng Comput, 29,* 159-166.

Liberati, D., Bedarida, L., Brandazza, P., & Cerutti, S. (1991b). A model for the cortico-cortical neural interaction in multisensory evoked potentials. *IEEE T Bio-Med Eng, 38*(9), 879-890.

Liberati, D., Cursi, M., Locatelli, T., Comi, G., & Cerutti, S. (1997). Total and partial coherence of spontaneous and evoked EEG by means of multi-variable autoregressive processing. *Med Biol Eng Comput, 35*(2), 124-130.

Muselli, M., & Liberati, D. (2000). Training digital circuits with hamming clustering. *IEEE T Circuits I, 47,* 513-527.

Muselli, M., & Liberati, D. (2002). Binary rule generation via hamming clustering. *IEEE T Knowl Data En, 14*(6), 1258-1268.

Pfurtscheller, G., Flotzinger, D., Mohl, W., & Peltoranta, M. (1992). Prediction of the side of hand movements from single-trial multi-channel EEG data using neural networks. *Electroenceph. Clin Neurophysiol, 82,* 313.

Roberts, S. J., & Penny, W. D. (2000). Real-time brain-computing interfacing: A preliminary study using Bayesian learning. *Medical & Biological Engineering & Computing, 38*(1), 56-61.

KEY TERMS

Artificial Neural Networks: Non-linear black box input-output relationships built on a regular structure of simple elements, loosely inspired to the natural neural system, even in learning by example.

Bayesian Statistics: Named after its developer, Bayes, it takes into account conditional probabilities in order to describe variable relationships.

Brain Computer Interfacing: Ability, even if only partial, of outside controlling by detecting intention through brain wave analysis.

Event-Related Potential: Electro-magnetic brain activity related to a specific event: it may be evoked from the outside, or self-ongoing.

Fuzzy Logic: Set theory mainly developed and made popular by Lofthi Zadeh at the University of California at Berkeley where belonging of elements is not crisply attributed to only one disjoint subset.

Parametric Identification: Black box mathematical modeling of an input-output relationship via simple, even linear equations depending on a few parameters, whose values do identify the system dynamics.

Piecewise Affine Identification: Linearization of a non-linear function, automatically partitioning data in subsets whose switching identifies state commuting in a hybrid dynamic-logical process.

Rule Induction: Inference from data of "if…then…else" rules describing the logical relationships among data.

Bridging Together Mobile and Service-Oriented Computing

Loreno Oliveira
Federal University of Campina Grande, Brazil

Emerson Loureiro
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

The growing popularity of powerful *mobile devices,* such as modern cellular phones, smart phones, and PDAs, is enabling *pervasive computing* (Weiser, 1991) as the new paradigm for creating and interacting with computational systems. Pervasive computing is characterized by the interaction of mobile devices with embedded devices dispersed across *smart spaces,* and with other mobile devices on behalf of users. The interaction between user devices and smart spaces occurs primarily through services advertised on those environments. For instance, airports may offer a notification service, where the system registers the user flight at the check-in and keeps the user informed, for example, by means of messages, about flight schedule or any other relevant information.

In the context of smart spaces, *service-oriented computing* (Papazoglou & Georgakopoulos, 2003), in short SOC, stands out as the effective choice for advertising services to mobile devices (Zhu, Mutka, & Ni, 2005; Bellur & Narendra, 2005). SOC is a computing paradigm that has in services the essential elements for building applications. SOC is designed and deployed through *service-oriented architectures* (SOAs) and their applications. SOAs address the flexibility for dynamic binding of services, which applications need to locate and execute a given operation in a pervasive computing environment. This feature is especially important due to the dynamics of smart spaces, where resources may exist anywhere and applications running on mobile clients must be able to find out and use them at runtime.

In this article, we discuss several issues on bridging mobile devices and service-oriented computing in the context of smart spaces. Since smart spaces make extensive use of services for interacting with personal mobile devices, they become the ideal scenario for discussing the issues for this integration. A brief introduction on SOC and SOA is also presented, as well as the main architectural approaches for creating SOC environments aimed at the use of resource-constrained mobile devices.

BACKGROUND

SOC is a distributed computing paradigm whose building blocks are distributed services. Services are self-contained software modules performing only pre-defined sets of tasks. SOC is implemented through the deployment of any software infrastructure that obeys its key features. Such features include loose coupling, implementation neutrality, and granularity, among others (Huhns & Singh, 2005). In this context, SOAs are software architectures complying with SOC features.

According to the basic model of SOA, service providers advertise service interfaces. Through such interfaces, providers hide from service clients the complexity behind using different and complex kinds of resources, such as databanks, specialized hardware (e.g., sensor networks), or even combinations of other services. Service providers announce their services in service registries. Clients can then query these registries about needed services. If the registry knows some provider of the required service, a reference for that provider is returned to the client, which uses this reference for contacting the service provider. Therefore, services must be described and published using some machine-understandable notation.

Different technologies may be used for conceiving SOAs such as grid services, Web services, and Jini, which follow the SOC concepts. Each SOA technology defines its own standard machineries for (1) service description, (2) message format, (3) message exchange protocol, and (4) service location.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

In the context of pervasive computing, services are the essential elements of smart spaces. Services are used for interacting with mobile devices and therefore delivering personalized services for people. Owning to the great benefits that arise with the SOC paradigm, such as interoperability, dynamic service discovery, and reusability, there is a strong and increasing interest in making mobile devices capable of providing and consuming services over wireless networks (Chen, Zhang, & Zhou, 2005; Kalasapur, Kumar, & Shirazi, 2006; Kilanioti, Sotiropoulou, & Hadjiefthymiades, 2005). The dynamic discovery and invocation of services are essential to mobile applications, where the user context may change dynamically, making different kinds of services, or service implementations, adequate at different moments and places.

However, bridging mobile devices and SOAs requires analysis of some design issues, along with the fixing of diverse problems related to using resources and protocols primarily aimed at wired use, as discussed in the next sections.

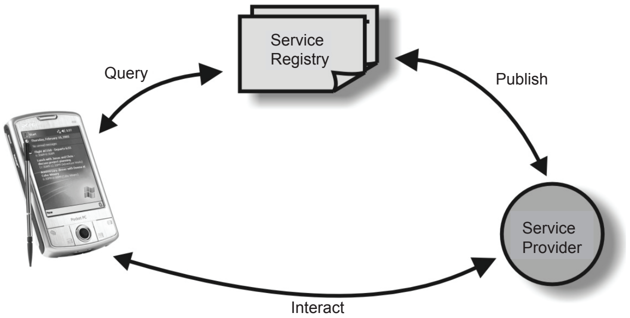
INTEGRATING MOBILE DEVICES AND SOAS

Devices may assume three different roles in a SOA: service provider, service consumer, or service registry. In what follows, we examine the most representative high-level scenarios of how mobile devices work in each situation.

Consuming Services

The idea is to make available, in a wired infrastructure, a set of services that can be discovered and used by mobile devices. In this context, different designs can be adopted for bridging mobile devices and service providers. Two major architectural configurations can be derived and adapted to different contexts (Duda, Aleksy, & Butter, 2005): direct communication and proxy aided communication. In Figure 1 we illustrate the use of direct communication.

Figure 1. Direct communication between mobile client and SOA infrastructure



In this approach, applications running at the devices directly contact service registries and service providers. This approach assumes the usage of fat clients with considerable processing, storage, and networking capabilities. This is necessary because mobile clients need to run applications coupled with SOA-defined protocols, which may not be suited for usage by resource-constrained devices.

However, most portable devices are rather resource-constrained devices. Thus, considering running on mobile devices applications with significant requirements of processing and memory footprint reduces the range of possible client devices. This issue leads us to the next approach, proxy-aided communication, illustrated in Figure 2.

In this architectural variation, a proxy is introduced between the mobile device and the SOA infrastructure, playing the role of mobile device proxy in the wired network. This proxy interacts via SOA-defined protocols with registries and service providers, and may perform a series of content adaptations, returning to mobile devices results using lightweight protocols and data formats.

This approach has several advantages over the previous one. The proxy may act as a cache, storing data of previous service invocations as well as any client relevant information, such as bookmarks and profiles. Proxies may also help client devices by transforming complex data into lightweight formats that could be rapidly delivered through wireless channels and processed by resource-constrained devices.

Advertising Services

In a general way, mobile devices have two choices for advertising services (Loureiro et al., 2006): the push-based approach and the pull-based approach. In the first one, illustrated in Figure 3, service providers periodically send the descriptions of the services to be advertised directly to potential clients, even if they are not interested in such services (1). Clients update local registries with information about available services (2), and if some service is needed, clients query their own registries about available providers (3).

In the pull-based approach, clients only receive the description of services when they require a service discovery. This process can be performed in two ways, either through centralized or distributed registries. In the former, illustrated in Figure 4, service descriptions are published in central servers (1), which maintain entries about available services (2). Clients then query this centralized registry in order to discover the services they need (3).

In the distributed registry approach, illustrated in Figure 5, the advertisement is performed in a registry contained in each provider (1). Therefore, once a client needs to discover a service, it will have to query all the available hosts in the environment (2) until discovering some service provider for the needed service (3).

Figure 2. Proxy intermediating communication between mobile client and SOA

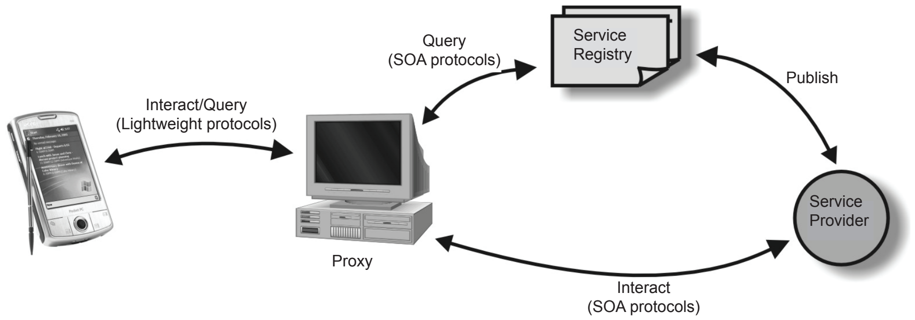

Query
(SOA protocols)
Service Registry
Publish
Interact/Query
(Lightweight protocols)
Service Provider
Proxy
Interact
(SOA protocols)

Figure 3. Push-based approach

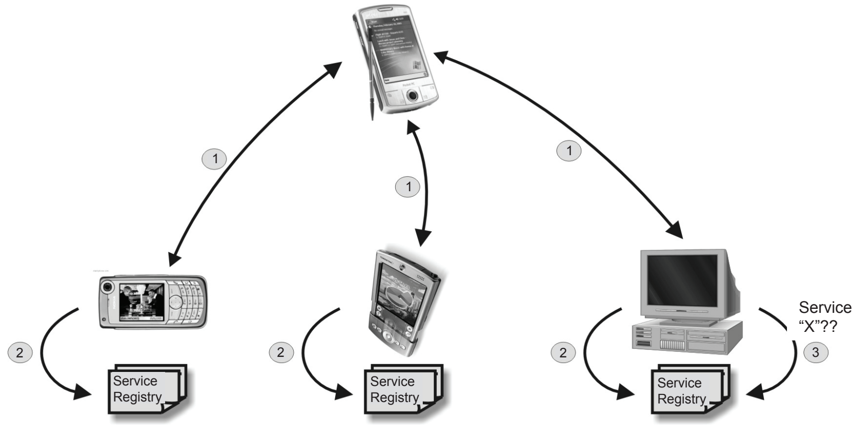
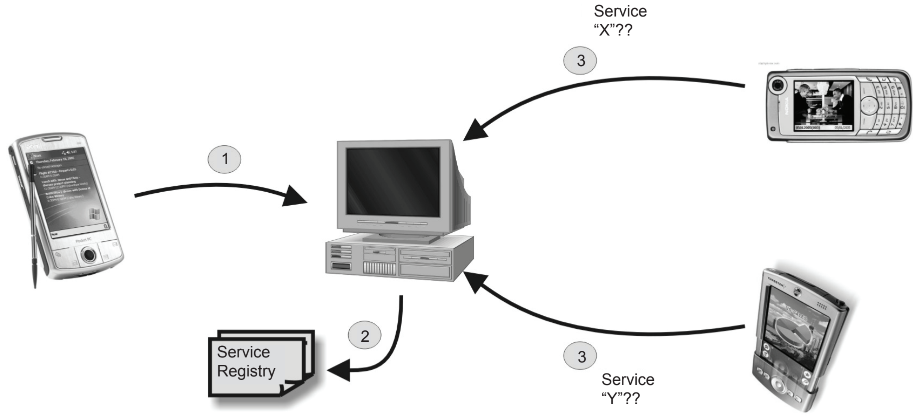

1
1
1
Service
"X"??
2
2
2
3
Service Registry
Service Registry
Service Registry

Figure 4. Pull-based approach with centralized registry

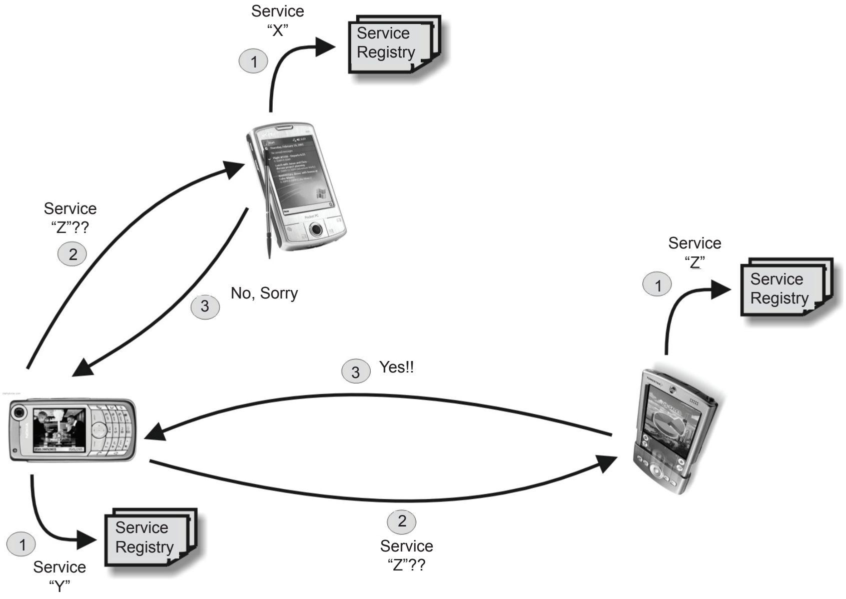

Service
"X"??
3
1
2
Service Registry
3
Service
"Y"??

ISSUES ON INTEGRATING MOBILE DEVICES AND SOAs

Regardless of using mobile devices for either consuming or advertising services in SOAs, both *mobility* and the *limitations* of these devices are raised as the major issues for this integration. Designing and deploying effective services aimed at mobile devices requires careful analysis of diverse issues related to this kind of service provisioning.

Next, we depict several issues that arise when dealing with mobile devices in SOAs. This list is not exhaustive, but rather representative of the dimension of parameters that should be balanced when designing services for mobile use.

Figure 5. Pull-based approach with distributed registries



Suitability of Protocols and Data Formats

SOAs are primarily targeted at wired infrastructures. Conversely, small mobile devices are known by their well-documented limitations. Thus, protocols and formats used in conventional SOAs may be inadequate for use with resource-constrained wireless devices (Pilioura, Tsalgatidou, & Hadjiefthymiades, 2002; Kilanioti et al., 2005).

For instance, UDDI and SOAP are, respectively, standard protocols for service discovery and messaging in Web services-based SOAs. When using UDDI for service discovery, multiple costly network round trips are needed. In the same manner, SOAP messages are too large and require considerable memory footprint and CPU power for being parsed. Hence, these two protocols impact directly in the autonomy of battery-enabled devices.

Disconnected and Connected Services

In the scope of smart spaces, where disconnections are the norm rather than the exception, we can identify two kinds of services (Chen et at., 2005): disconnected and connected services. The first ones execute by caching the inputs of users in the local device. Once network connectivity is detected, the service performs some sort of synchronization. Services for messaging (e.g., e-mail and instant messages) and field research (e.g., gathering of data related to the selling of a specific product in different supermarkets) are some examples of services that can be implemented as disconnected ones.

Connected services, on the other hand, are those that can only execute when network connectivity is available in the device. Some examples of connected services include price checking, ordering, and shipment tracking. Note, however, that these services could certainly be implemented as disconnected services, although their users will generally need the information when demanded, neither before nor later. Therefore, there is no precise categorization of what kind of services would be connected or disconnected, as this decision is made by the system designer.

User Interface

User interfaces of small portable devices are rather limited in terms of screen size/resolution and input devices, normally touch screens or small built-in keyboards. This characteristic favors services that require low interaction to complete transactions (Pilioura et al., 2002). Services requiring many steps of data input, such as long forms, tend to: stress users, due to the use of non-comfortable input devices; reduce device autonomy, due to the extra time for typing data; and increase the cost of data transfer, due to larger amounts of data being transferred.

A possible alternative for reducing data typing by clients is the use of context-aware services (Patterson, Muntz, & Pancake, 2003). Context-aware services may reduce data input operations of mobile devices by inferring, or gathering through sensors, information about a user's current state and needs.

Frequent Temporary Disconnections

Temporary disconnections between mobile device and service provider are common due to user mobility. Thus, both client applications and service implementations must consider the design of mechanisms for dealing with frequent disconnections.

Different kinds of services require distinct solutions for dealing with disconnections. For instance, e-business applications need machineries for controlling state of transactions and data synchronization between mobile devices and service providers (Sairamesh, Goh, Stanoi, Padmanabhan, & Li, 2004). Conversely, streaming service requires seamless reestablishment and transference of sessions between access points as the user moves (Cui, Nahrstedt, & Xu, 2004).

Security and Privacy

Normally, mobile devices are not shared among different users. Enterprises may benefit from this characteristic for authenticating employees, for instance. That is, the system knows the user and his/her access and execution rights based on profiles stored in his/her mobile devices. However, in commercial applications targeted at a large number of unknown users, this generates a need of anonymity and privacy of consumers. This authentication process could cause problems, for example in case of device thefts, because the device is authenticated and not the user (Tatli, Stegemann, & Lucks, 2005).

Security also has special relevance when coping with wireless networks (Grosche & Knospe, 2002). When using wireless interfaces for information exchange, mobile devices allow any device in range, equipped with the same wireless technology, to receive the transferred data. At application layer, service providers must protect themselves from opening the system to untrusted clients, while clients must protect themselves from exchanging personal information with service providers that can use user data for purposes different than the ones implicit in the service definition.

Device Heterogeneity and Content Adaptation

Modern mobile devices are quite different in terms of display sizes, resolutions, and color capabilities. This requires services to offer data suitable for the display of different sorts of devices. Mobile devices also differ in terms of processing capabilities and wireless technologies, which makes harder the task of releasing adequate data and helper applications to quite different devices.

Therefore, platform-neutral data formats stand out as the ideal choice for serving heterogeneous sets of client devices. Another possible approach consists of using on-demand data adaptation. Service providers may store only one kind of best-suited data format and transform the data, for example, using a computational grid (Hingne, Joshi, Finin, Kargupta, & Houstis, 2003), when necessary to transfer the data to client devices. Moreover, dynamic changes of conditions may also require dynamic content adaptation in order to maintain pre-defined QoS threshold values. For instance, users watching streamed video may prefer to dynamically reduce video quality due to temporary network congestion, therefore adapting video data, and to maintain a continuous playback instead of maintaining quality and experiencing constant playback freezing (Cui et al., 2004).

Consuming Services

As discussed before, system architects can choose between two major approaches for accessing services of SOAs from mobile devices: direct communication and proxy-aided communication. The two approaches have some features and limitations that should be addressed in order to deploy functional services. Direct communication suffers from the limitations of mobile devices and relates to other discussions presented in this article, such as adequacy of protocols and data formats for mobile devices and user interface.

If, on the one hand, proxy-aided communication seems to be the solution for problems of the previous approach, on the other hand it also brings its own issues. Probably the most noted is that proxies are single points of failures.

Furthermore, some challenges related to wired SOAs are also applicable to both approaches discussed. Service discovery and execution need to be automated to bring transparency and pervasiveness to the service usage. Moreover, especially in the context of smart spaces, services need to be personalized according to the current user profile, needs, and context. Achieving this goal may require describing the semantics of services, as well as modeling and capturing the context of the user (Chen, Finin, & Joshi, 2003).

Advertising Services

A number of issues and technical challenges are associated with this scenario. The push-based approach tends to consume a lot of bandwidth of wireless links according to the number of devices in range, which implies a bigger burden over mobile devices.

Using centralized registries creates a single point of failure. If the registry becomes unreachable, it will not be possible to advertise and discover services. In the same manner, the discovery process is the main problem with the approach of distributed registries, as it needs to be well designed in order to allow clients to query all the hosts in the environment.

Regardless of using centralized or distributed registries, another issue rises with mobility of service providers. When service providers move between access points, a new address is obtained. This changing of address makes service providers inaccessible by clients that query the registry where it published its services. Mechanisms for updating the registry references must be provided in order for services to continue to be offered to their requestors.

FUTURE TRENDS

The broad list of issues presented in this article gives suggestions about future directions for integrating SOC and mobile devices. Each item depicted in the previous section is already an area of intensive research. Despite this, both SOC and mobile computing still lack really functional and mature solutions for the problems presented.

In particular, the fields of context-aware services and security stand out as present and future hot research fields. Besides, the evolution itself of mobile devices towards instruments with improved processing and networking power, as well as better user interfaces, will reduce the complexity of diverse challenges presented in this article.

CONCLUSION

In this article we have discussed several issues related to the integration of mobile devices and SOC. We have presented the most representative architectural designs for integrating mobile devices to SOAs, both as service providers and service consumers.

While providing means for effective integration of mobile devices and service providers, SOC has been leveraging fields such as mobile commerce and pervasive computing. Nonetheless, several issues remain open, requiring extra efforts for designing and deploying truly functional services.

REFERENCES

Bellur, U., & Narendra, N. C. (2005). Towards service orientation in pervasive computing systems. *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)* (Vol. II, pp. 289-295).

Chen, H., Finin, T., & Joshi, A. (2003). An ontology for context-aware pervasive computing environments. *The Knowledge Engineering Review, 18*(3), 197-207.

Chen, M., Zhang, D., & Zhou, L. (2005). Providing Web services to mobile users: The architecture design of an m-service portal. *International Journal of Mobile Communications, 3*(1), 1-18.

Cui, Y., Nahrstedt, K., & Xu, D. (2004). Seamless user-level handoff in ubiquitous multimedia service delivery. *Multimedia Tools Applications, 22*(2), 137-170.

Duda, I., Aleksy, M., & Butter, T. (2005). Architectures for mobile device integration into service-oriented architectures. *Proceedings of the 4th International Conference on Mobile Business (ICBM'05)* (pp. 193-198).

Grosche, S.S., & Knospe, H. (2002). Secure mobile commerce. *Electronics & Communication Engineering Journal, 14*(5), 228-238.

Hingne, V., Joshi, A., Finin, T., Kargupta, H., & Houstis, E. (2003). Towards a pervasive grid. *Proceedings of the 17th International Parallel and Distributed Processing Symposium (IPDPS'03)* (p. 207.2).

Huhns, M.N., & Singh, M.P. (2005). Service-oriented computing: Key concepts and principles. *IEEE Internet Computing, 9*(1), 75-81.

Kalasapur, S., Kumar, M., & Shirazi, B. (2006). Evaluating service oriented architectures (SOA) in pervasive computing. *Proceedings of the 4th IEEE International Conference on Pervasive Computing and Communications (PERCOMP'06)* (pp. 276-285).

Kilanioti, I., Sotiropoulou, G., & Hadjiefthymiades, S. (2005). A client/intercept based system for optimized wireless access to Web services. *Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05)* (pp. 101-105).

Loureiro, E., Bublitz, F., Oliveira, L., Barbosa, N., Perkusich, A., Almeida, H., & Ferreira, G. (2007). Service provision for pervasive computing environments. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce.* Hershey, PA: Idea Group Reference.

Papazoglou, M. P., & Georgakopoulos, D. (2003). Service-oriented computing: Introduction. *Communications of the ACM, 46*(10), 24-28.

Patterson, C. A., Muntz, R. R., & Pancake, C. M. (2003). Challenges in location-aware computing. *IEEE Pervasive Computing, 2*(2), 80-89.

Pilioura, T., Tsalgatidou, A., & Hadjiefthymiades, S. (2002). Scenarios of using Web services in m-commerce. *ACM SIGecom Exchanges, 3*(4), 28-36.

Sairamesh, J., Goh, S., Stanoi, I., Padmanabhan, S., & Li, C. S. (2004). Disconnected processes, mechanisms and architecture for mobile e-business. *Mobile Networks and Applications, 9*(6), 651-662.

Tatli, E. I., Stegemann, D., & Lucks, S. (2005). Security challenges of location-aware mobile business. *Proceedings* *of the 2nd IEEE International Workshop on Mobile Commerce and Services (WMCS'05)* (pp. 84-95).

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 66-75.

Zhu, F., Mutka, M. W., & Ni, L. M. (2005). Service discovery in pervasive computing environments. *IEEE Pervasive Computing, 4*(4), 81-90.

KEY TERMS

Grid Service: A kind of Web service. Grid services extend the notion of Web services through the adding of concepts such as statefull services.

Jini: Java-based technology for implementing SOAs. Jini provides an infrastructure for delivering services in a network

Mobile Device: Any low-sized portable device used to interact with other mobile devices and resources from smart spaces. Examples of mobile devices are cellular phones, smart phones, PDAs, notebooks, and tablet PCs.

Proxy: A network entity that acts on behalf of another entity. A proxy's role varies since data relays to the provision of value-added services, such as on-demand data adaptation.

Streaming Service: One of a number of services that transmit some sort of real-time data flow. Examples of streaming services include audio streaming or digital video broadcast (DVB).

Web Service: Popular technology for implementing SOAs built over Web technologies, such as XML, SOAP, and HTTP.

Browser-Less Surfing and Mobile Internet Access

Gregory John Fleet
University of New Brunswick at Saint John, Canada

Jeffery G. Reid
xwave Saint John, Canada

INTRODUCTION

Lately, we have seen the use of a number of new technologies (such as Javascript, XML, and RSS) used to show how Web content can be delivered to users without a traditional browser application (e.g., Microsoft Explorer). In parallel, a growing number of PC applications, whose main job previously was to manage local resources, now are adding Internet connectivity to enhance their role and use (e.g., while iTunes started as a media player for playing and managing compressed audio files, it now includes Web access to download and purchase music, video, podcasts, television shows, and movies).

While most attempts at providing Internet access on mobile devices (whether wireless phones or personal digital assistants) have sought to bring the traditional browser, or a mobile version of the browser, to these smaller devices, they have been far from successful (and a far cry from the richer experience provided by browsers on the PC using standard input and control devices of keyboards and a mouse). Next, we will highlight a number of recent trends to show how these physical and use-case constraints can be significantly diminished.

BACKGROUND

Mobile telephony and mobile computing continue to display unprecedented growth worldwide. Zee News (2005) reports that in some parts of the world, such as India, mobile phones are now more popular than traditional landline phones. Since 2000, many developed countries have spent large amounts of money on the installation and deployment of wireless communication infrastructure (Kunz & Gaddah, 2005). And this growth trend is not confined to the mobile phone handset market. It is also being experienced across other mobile devices. In fact, in 2004 more mobile phones shipped than both automobiles and personal computers (PCs) combined, making them the fastest adopted consumer product of all time (Clarke & Flaherty, 2005). Further, Wiberg (2005) points out that this increase in mobile device usage spans across business and non-business usage. Therefore, this growth is not simply due to increased consumer demand; businesses are continually seeing new value in equipping employees with mobile computing and communication devices.

There has also been steady growth in the use of the Internet, as well as in the nature of Internet usage. The size of the Internet, measured in terms of the number of users, is more than 800 million users (Global Reach, 2004). While the majority of the users are English, other languages are experiencing significant growth in the number of users, and this growth is expected to continue, given the large numbers of non-English-speaking populations.

Some of the drivers for the increase in Internet usage include the growth in Web-enabled applications and the availability of high-speed, always-on Internet (Bink, 2004).

Kunz and Gaddah (2005) identify two broad technological developments that are converging to enable mobile computing (the use of the Internet through mobile devices). The first of these technological developments is the accessibility to the Internet regardless of location, as evidenced by the growth in wireless *hotspots*. Now users can connect to the Internet from various locations and access Internet content without being connected to a physical local area network (LAN) connection or other type of landline connection.

The second technological development is the drive to reduce the size of computer hardware (Kunz & Gaddah, 2005). This size reduction increases the portability of these devices, leading to the mobile nature of the devices as well as the desire to connect these devices to the Internet.

Unfortunately, being *able* to provide Internet access to mobile devices has not *ensured* a quality Web experience. The next section will profile the current mobile Web experience.

USER EXPERIENCE OF WEB ON MOBILE DEVICES

The Web Browser on a PC

Let us start with the typical experience of the Web. The most common way to navigate the Internet is through the

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

use of a browser, a software application that allows the user to locate and display Web pages (Webopedia, 2006). On the personal computer (PC), there are a variety of browsers available, including Microsoft Internet Explorer, Mozilla Firefox, Opera, Netscape, Apple Safari, and Konqueror (Wikipedia, 2006a).

A cross-section of definitions from the Web outlines the basic functionality of these browsers (http://www.google.com/search?hl=en&lr=&q=define%3A+web+browser&btnG=Search); the Web browser is a graphical interface (i.e., icons, buttons, menu options) that:

- interprets HTML files (resources, services) from Web servers, and formats them into Web pages; and
- provides the ability to both view and interact with Web content (including download and upload of media content).

Yet most modern browsers also include additional functionality, assisting with the management of the tool's functionality and the content to which they provide access. This functionality includes:

- **Bookmarking:** The ability to save and manage Web addresses.
- **Cookies and Form-Filling:** The ability of the browser to pre-fill form fields (e.g., address or contact information), or provide the Web server with identifying information in order to customize the content received from the server.
- **Searching:** The ability to conduct a Web or local file search.
- **History:** The automatic cataloging of previously visited Web sites.
- **Display Modification:** The ability to customize the way Web content is displayed (e.g., size of text, types of media files that can be viewed, etc.)

It is also important to note that in the typical use of a Web browser, the user searches for information on the Web, often starting with a broad search and successively narrowing that search to meet his or her information goal (i.e., to go from the general to the specific).

The Web Browser on a Mobile Device

For mobile devices, such as cell phones or personal digital assistants (PDAs), the Web browser application is often referred to as a microbrowser (also minibrowser and mobile browser; see Wikipedia, 2006b). The difference between a *full* browser and the microbrowser is that the code in the microbrowser application has been optimized to accommodate the smaller screens, memory, and bandwidth limitations of mobile devices. In addition, the Web servers often communicate with these microbrowsers using variations on the standard HTML (hypertext markup language), again to accommodate the screen, memory, and bandwidth restrictions.

Internet usage on mobile devices poses a number of challenges that are different than those found on a traditional computing device such as a PC (Becker, 2005). As mentioned previously, the computing power (processor and memory configuration), the transmission bandwidth, and screen size on the mobile device are really just a fraction of what users have available to them on a PC. More importantly, the limitations in screen size and physical interface often require users to restrict the activities they might otherwise seek to accomplish on the Web.

The physical restrictions (that being the telephone keypad and four-way scroll and navigation keys) can be quite significant. On a PC, we have a full-sized QWERTY keyboard and mouse interface for entering searches and addresses, or navigating Web pages. On the mobile device, in particular the cell phone, these physical input and control devices are replaced with a keypad designed for dialing phone numbers (not entering text strings), and horizontal/vertical navigation keys that significantly slow simple scrolling and selection of content.[1] In user studies, Chen, Xie, Ma, and Zhang (2005) report that users, when browsing the Web on a phone, handheld computer, or personal digital assistant, spend the majority of their time scrolling the screen to locate and select the content of interest.

Despite these real challenges, Nugent (2005) expects that the need for mobile Web browsing will increase, and people will want these devices to stay small, weigh less, cost less, run cooler and longer on one charge, but continue to do more than today's devices. Lawton (2001) believes that meeting these needs will require faster wireless connections, larger displays, as well as new usage paradigms and/or content that fits these smaller devices.

This is the environment mobile users are operating in today. A user can either struggle with a small screen and content that does not fit within that screen, or lug around a larger device that has an adequately sized screen but more limited connectivity options.

Technology and Service Barriers

There are a number of technological hurdles that need to be overcome for widespread adoption of mobile Internet usage. Chan and Fang (2005) identify a number of technological barriers, which range from connectivity and bandwidth issues to the lack of standards and broad use of proprietary tools and languages. Kuniavsky (2006) also notes the numerous and often complex relationships that exist between the multiple service, application, and technology providers currently needed to deliver mobile computing to the user, and how none of these players is wholly responsible for the resulting user experience.

Identification of the challenges associated with mobile Internet usage is only part of the problem. After the issues have been identified, developing solutions to deal with the problems is the next step in the process. Many manufacturers are presently making moves to deal with some of these identified issues. One common move is to increase the screen size of the mobile device. One example of this trend is the Sony Ericsson P800 SmartPhone, which has increased the size of its display area while attempting to maintain the overall size of the device itself. Another more recent example is the Sony Ericsson P910i, with its larger screen, miniature QWERTY keyboard, and pen-based interface.

Another design approach to deal with limited screen space is to focus on the content rather than the size of the screen, as is attempted with standards such as WAP, WML, HDML, as well as services such as i-mode (Chen et al., 2005).

What is needed, though, is the development of Web content and mobile applications that can be viewed, navigated, and controlled from small devices (Nayak, 2005), because, at this time, consumers find the small screen display and small buttons on these devices difficult to use. Chan and Fang (2005) believe that these technologies need to mature, and until that time, the mobile Internet will be geared toward applications requiring limited bandwidth, short exchange of data and text, and simple functionality. Therefore, using smaller mobile devices to perform tasks similar to those carried out on a traditional computing platform poses challenges for users and manufacturers alike.

FUTURE TRENDS

Interestingly, there are a number of new technologies and trends that might suggest an evolution of the mobile computing Web experience. This evolution comes from a number of different places. In this section of the article, a few specific trends are highlighted in order to demonstrate this potential.

Web-Enabled Desktop Clients

In recent years, we are seeing more and more desktop clients (or applications) reaching beyond the processes of the PC and the content on the local drive to networked and Internet resources and content. Apple's iTunes media player was first released as a desktop application for playing music files from one's hard drive. Since that first release, it has grown to not only allow streaming of music libraries over networks, but now has a built-in Web browser tied to one of the most successful online music stores today.

There are additional examples of traditional desktop clients that have added Web connectivity to their functional specifications; these include desktop applications with built-in version checking; address book applications that communicate with LDAP servers; Google desktop™, providing the ability to search and find information not only on your local drives, but also on e-mail and Web servers; and cataloguing programs that match your own library of CDs, books, or DVDs with online databases such as Internet Movie Database (imd.com) or Amazon.com. I suspect this trend is only just beginning, and we will continue to see additional examples as software companies add both Internet connectivity and imbedded browsers into desktop applications in order to add new and unique value for users.

Webtop Clients

At the same time, there are also some exciting examples of Web applications (or services) that only require a standard Web browser. Web services have been around since the beginning of the Web, but what differentiates these newer Web applications is their attention to usability and responsiveness, resulting in a *Webtop client* that responds and behaves in ways similar to a desktop client. For example, Flickr.com allows individuals to upload photos from their cameras and hard drives to the Flickr Web servers. Then, in desktop-client fashion, they allow us to arrange the order with simple drag-and-drop, or name, edit, and tag photo labels by directly clicking on the titles within the Web browser.

Other examples include the Web services of Google, MyYahoo, and MyMSN, as well as the excellent services from 37signals.com (Basecamp, Backpack, Writeboard, Ta-Da List, and the growing number of services built using the Ruby on Rails development environment). In all these examples, the responsiveness of these Web clients is quite impressive, mimicking the behavior of their desktop counterparts.

The Changing Mental Model of Web Access

These examples of desktop and Webtop clients demonstrate a blurring of the lines between Web browsing or surfing and running local applications. I believe this is a good thing, since it suggests that Web connectivity is not limited to what is accomplished and viewed through a browser. And for mobile devices, this lack of distinction should also be a good thing—allowing users to think about Internet content separate from traditional browsers. This could also suggest that users might adapt their user model of expecting Internet content (especially on small devices) to be only through an Internet browser.

Seeing the Internet on mobile devices as separate from a browser is only one (significant) step in producing a better user experience. It is also important to recognize the other constraints that limit a quality Internet experience on these small devices: the constrained visual and physical interface.

Reproducing Web sites onto small screens, at best, requires the ability to visualize content beyond the screen, and, at worst, produces a frustrating, unacceptable experience.

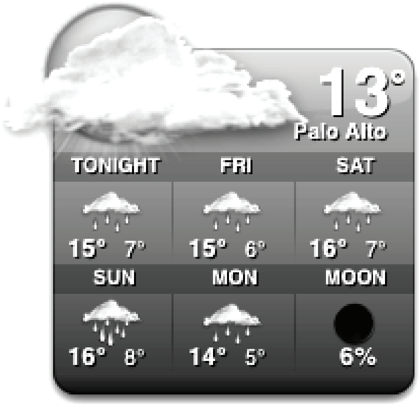
SOLVING THE VISUAL LIMITS OF MOBILE DEVICES

A variety of technologies (XML, ATOM, Javascript, WebKit) have been used of late to create a number of useful Web services. One of the most common is RSS feeds, where the user can *subscribe* to the content found on a Web server. The current implementation of RSS satisfies two user goals: to filter Web content to only those topics of interest, and to provide real-time notification of updates to the Web site. Therefore, RSS provides a technology to allow users to *browse* the Web in a more focused manner, providing personalized views of self-selected content.

Another example of viewing self-selected Web content is found with Yahoo's Konfabulator (also known as the Yahoo! Widget Engine—see http://widgets.yahoo.com). This desktop client is a real-time Javascript compiler that can execute small Javascript files (called *widgets*) to accomplish whatever task they have been programmed to accomplish. The result is a small, windowless, (and in the case of scripts that communicate with Web servers) browser-less view of live Web content. Figure 1 shows an example using the Weather widget, which displays live weather conditions and the five-day forecast for a particular location configured by the user. Additional Web information is available with mouse-over or clicks on the widget.

The latest operating system from Apple (known as Tiger or 10.4) has also added similar functionality for displaying self-selected Web content. More appropriately named, these Dashboard™ widgets organize and/or present Web content in a way that is easy to read. Presently, thousands of widgets are available for download, whether from Yahoo's Web site or Apple.com (as of January 2006, there were more than 4,000 widgets available for download). What is interesting about widgets is the fact that most are designed to present their Web content in a fairly constrained visual space, separate from large resource requirements or visual real estate needs. In other words, these widgets provide what could be a perfect example of self-selected, rich Internet content for small screens.

Figure 1. The Weather widget, using Yahoo's Widget Engine, showing current and forecasted weather for Palo Alto, California



Another interesting example comes from some software developers in Japan. They have demonstrated the ability to create a Dashboard™ widget of a full Web browser, only miniaturized to dimensions that could easily work on a standard PDA-size screen (see http://hmdt-web.net/shiira/mini/en). Therefore, with a high-quality display such as that found on today's mobile devices and iPods, it is quite easy to imagine using this miniaturized view of Web pages to surf the Web, especially on Web sites where the format and layout is familiar.

Now we turn to the problem of the physical (input and control) interface on mobile devices. If you have this view of a miniaturized Web page, how do you move around and select the buttons and links on the page? Using a four-way scroll key might work for the limited content in most Widgets, but is a very poor substitute for a keyboard and mouse when browsing a full (though miniaturized) Web page.

Solving the Physical Interface of Mobile Devices

A keyboard and mouse is not just the standard input and control device for Web surfing, but provides a rich interaction for control and text input. A telephone keypad and four-way scroll key does not even come close to that user experience, therefore making the Web experience on mobile devices very constrained indeed.

Yet we have an excellent example of one specific mobile device that has, over the past four years, shown that navigation of large hierarchies of data can be very quickly accomplished with only a finger or thumb. The iPod music player provides both a simple and extremely intuitive interface for moving through and selecting from vast playlists, photos, folders, and files—using the scroll wheel or click wheel[2] design. The click wheel interface is currently only available on Apple's iPod music and video players, but there has been much discussion of the possibility of this interface being used on other small devices, such as cell phones or PDAs (see Shortflip, 2006; Baig, 2005). After experiencing how easy it is to use an iPod to navigate and select music, photos, or videos, it is not difficult to imagine the same physical interface being used to select phone numbers from an address book, and even select and navigate content on a Web page.

More recently, there has been discussion of some patents filed at the U.S. Patent Office site (see http://tinyurl.com/8zxuv). The patent application document demonstrates a device where the whole front is a display screen, and the control interface comes from a touch-activated, touch-sensitive click wheel that is available from the visual display wherever the user touches the screen. Therefore, the combination of a large screen, with a touch-activated telephone keypad and/or click wheel, provides a compelling possibility for a high-fidelity Web experience on a mobile device.

CONCLUSION

In this article, we have argued that today's mobile devices provide a poor user experience when presenting Internet and Web content. These devices have two major physical constraints: a visual display that is too small to present the typical browser-based view of the Web, and a physical interface (telephone keypad and four-way scroll key with center select) that is not easily adapted to text entry or interface control. Users, therefore, are unable to reproduce the familiar encyclopedic browsing of Web content on these miniature visual and physical interfaces.

At the same time, a number of new trends has been highlighted, demonstrating the possibility of producing a much stronger and more compelling user experience of the Web on small mobile devices.

REFERENCES

Baig, E. C. (2005). New iTunes phone a snazzy device. *USA Today.com.* Retrieved February 1, 2006, from http://www.usatoday.com/tech/columnist/edwardbaig/2005-09-07-itunes-phone_x.htm

Becker, S. A. (2005). Web usability. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 5, pp. 3074-3078). Hershey, PA: Idea Group Reference.

Bink, S. (2004). Browserless Net use on the rise. *Bink.nu.* Retrieved February 1, 2006, from http://bink.nu/Article798.bink

Chan, S. S., & Fang, X. (2005). Interface design issues for mobile commerce. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 3, pp. 1612-1616). Hershey, PA: Idea Group Reference.

Chen, Y., Xie, X., Ma, W., & Zhang, H. (2005). Adapting Web pages for small-screen devices. *IEEE Internet Computing.* Retrieved February 1, 2006, from http://research.microsoft.com/~xingx/tic1.pdf

Clarke, I., & Flaherty, T. (2005). Portable portals for m-commerce. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 4, pp. 2293-2296). Hershey, PA: Idea Group Reference.

Fleet, G. J. (2003, October 16-18). The devolution of the Web browser: The fracturing of Internet Explorer. *Proceedings of the Atlantic Schools of Business Conference,* Halifax, Nova Scotia, Canada.

Global Reach. (2004). *Global Internet statistics by language.* Retrieved February 1, 2006, from http://global-reach.biz/globstats/index.php3

Kuniavsky, M. (2006). User experience and HCI. In J. Jacko & A. Sears (Eds.), *The human-computer interaction handbook* (2nd ed.). Retrieved February 1, 2006, from http://www.orangecone.com/hci_UX_chapter_0.7a.pdf

Kunz, T., & Gaddah, A. (2005). Adaptive mobile applications. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 1, pp. 47-52). Hershey, PA: Idea Group Reference.

Lawton, G. (2001). Browsing the mobile Internet. *IEEE Computer, 35*(12), 18-21.

Nugent, J. H. (2005). Critical trends in telecommunications. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 1, pp. 634-639). Hershey, PA: Idea Group Reference.

Nayak, R. (2005). Wireless technologies to enable electronic business. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 5, pp. 3101-3105). Hershey, PA: Idea Group Reference.

Shortflip.com. (2006). *The future of the iPod.* Retrieved February 1, 2006, from http://www.shortflip.com/article/The-Future-of-the-iPod-149.html

Webopedia. (2004). *Browser.* Accessed February 1, 2006, from http://www.webopedia.com/TERM/B/browser.html

Wiberg, M. (2005). "Anytime, anywhere" in the context of mobile work. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 1, pp. 131-134). Hershey, PA: Idea Group Reference.

Wikipedia. (2006a). *Web browser.* Accessed February 1, 2006, from http://en.wikipedia.org/wiki/Web_browser

Wikipedia. (2006b). *Microbrowser.* Accessed February 1, 2006, from http://en.wikipedia.org/wiki/Microbrowser

Zee News. (2005). *Mobile phones outpace landline but with grey calls.* Retrieved February 1, 2006, from http://www.zeenews.com/znnew/articles.asp?aid=194056&sid=ZNS

KEY TERMS

Atom: One of a number of Web formats that supports user subscription to online content.

Click Wheel: The physical interface on Apple's iPod for moving through the directories and selecting items.

HDML: Handheld Device Markup Language.

i-mode: A popular wireless Internet service initially available only in Japan.

Javascript: Scripting programming language.

LDAP: Lightweight directory access protocol.

Podcast: The distribution of audio or video content over the Web using Atom or RSS.

RSS: Rich site summary or really simple syndication.

Ruby on Rails: An new open-source Web application framework.

WAP: Wireless application protocol.

WebKit: Application framework for Apple's Safari Web browser.

WML: Wireless Markup Language.

XML: eXtensible Markup Language.

ENDNOTES

[1] It is true that some PDAs and cell phones are using miniature QWERTY keyboards for an input device, though the tiny size is only marginally better than the keypad.

[2] The click wheel interface allows the user to navigate a vertical array of items or folders by rotating the wheel either clockwise or counterclockwise. Selecting an item in the list or moving deeper into the folder structures can be accomplished with the center button or the four buttons placed 90 degrees apart.

Building Web Services in P2P Networks

Jihong Guan
Tongji University, China

Shuigeng Zhou
Fudan University, China

Jiaogen Zhou
Wuhan University, China

INTRODUCTION

Nowadays peer-to-peer (P2P) and Web services are two of the hottest research topics in computing. Roughly, they appear as two extremes of distributed computing paradigm. Conceptually, P2P refers to a class of systems and applications that employ distributed resources to perform a critical function in a decentralized way. A P2P distributed system typically consists of a large number of nodes (e.g., PCs connected to the Internet) that can potentially be pooled together to share their resources, information, and services. These nodes, taking the roles of both consumer and provider of data and/or services, may join and depart the P2P network at any time, resulting in a truly dynamic and ad-hoc environment. Apart from improving scalability by avoiding dependency on centralized servers, the distributed nature of such a design can eliminate the need for costly infrastructure by enabling direct communication among clients, along with enabling resource aggregation, thus providing promising opportunities for novel applications to be developed (Ooi, Tan, Lu, & Zhou, 2002).

On the other hand, Web services technologies provide a language-neutral and platform-independent programming model that can accelerate application integration inside and outside the enterprise (Gottschalk, Graham, Kreger, & Snell, 2002). It is convenient to construct flexible and loosely coupled business systems by application integration under a Web services framework. Considering Web services are easily applied as wrapping technology around existing applications and information technology assets, new solutions can be deployed quickly and recomposed to address new opportunities. With the acceleration of Web services adoption, the pool of services will grow, fostering development of more dynamic models of just-in-time application and business integration over the Internet.

However, current proposals for Web services infrastructures are mainly based on centralized approaches such as UDDI: a central repository is used to store services descriptions, which will be queried to discover or, in a later stage, compose services. Such centralized architecture is prone to introducing single points of failure and hotspots in the network, and exposing vulnerability to malicious attacks. Furthermore, making full use of Web services capabilities using a centralized system does not scale gracefully to a large number of services and users. This difficulty is severe by the evolving trend to ubiquitous computing in which more and more devices and entities become services, and service networks become extremely dynamic due to constantly arriving and leaving service providers.

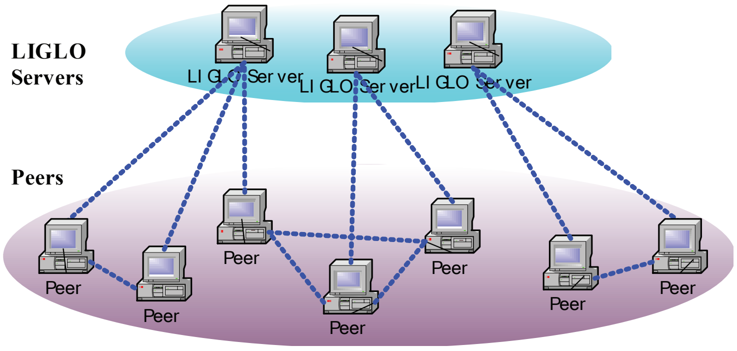
We explore the techniques of building Web services systems in a P2P environment. By fitting Web services into a P2P environment, we aim to add more flexibility and autonomy to Web services systems, and alleviate to some degree the inherent limitations of these centralized systems. As a case study, we present our project *BP-Services.* BP-Services is an experimental Web services platform built on BestPeer (http://xena1.ddns.comp.nus.edu.sg/p2p/)—a generic P2P infrastructure designed and implemented collaboratively by the National University of Singapore and Fudan University of China (Ng, Ooi, & Tan, 2002).

FITTING WEB SERVICES INTO A P2P FRAMEWORK

A *Web service* can be seen as an interface that describes a collection of operations that are network accessible through standardized XML messaging (Gottschalk et al., 2002). Web services consist of three roles and three operations: the roles are *providers, requesters,* and *registrars* of services, and the operations are *publish, find,* and *bind.* The service providers are responsible for creating Web services and corresponding service definitions, and then publishing the services with a service registry based on UDDI specification. The service requesters first find the services requested via the UDDI interface, and the UDDI registry provides the requesters with WSDL service descriptions and URLs pointing to these services themselves. With the information obtained, the requesters can then bind directly with the services and *invoke* them.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Network topology of BestPeer



Over the last few years, many P2P systems have been developed and deployed for different purposes and with different technologies, such as Napster (http://www.napster.com/), Gnutella (http://gnutella.wego.com/), and Freenet (http://freenet.sourceforge.com/), to name a few. The architecture of these systems can be categorized into three groups mainly based on their network topologies: centralized P2P, pure P2P, and hybrid P2P systems (Yang & Garcia-Molina, 2001). In a centralized P2P network, there is a central server responsible for maintaining indexes on the metadata for all the peers in the network. Pure P2P is simply P2P systems with fully autonomous peers—that is, all nodes are equal, no functionality is centralized, and the communication between peers is also symmetric. Hybrid P2P is a kind of tradeoff between centralized P2P and pure P2P, which is structured hierarchically with a supernode layer and a normal peers layer.

Fitting Web services into P2P framework is to adapt Web services to P2P environment, which results in the so-called *P2P Web services,* or simply *P2P services*. Here P2P service is different from the ordinary Web services at least in three aspects. First, typically a peer in P2P services takes all three roles of services provider, consumer, and registrar, whereas in ordinary Web services, a node can typically be a producer and/or a consumer, but not a registrar at the same time. Second, generally speaking, servers in ordinary Web services systems are well-known hosts, with static IP addresses and on the outside of a firewall. However, this is not usually the case in the P2P world. A services node may join or depart the P2P services network at any time. Third, the preferred method of finding Web services in the ordinary architecture is currently through a central repository known as a UDDI operator. Nevertheless, P2P services systems have no central server to hold UDDI registry; each peer node manages its own UDDI registry locally. So, new and efficient mechanisms for services discovery in P2P services environment are required.

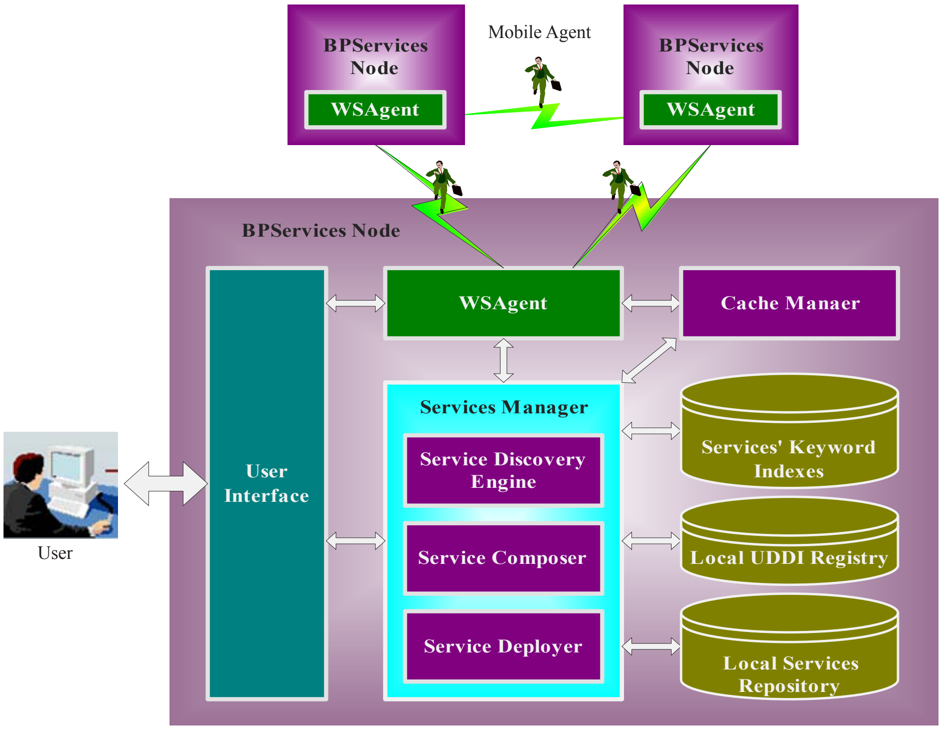
Corresponding to the architecture of P2P systems, there may also be three schemes for building P2P services applications: centralized P2P services, pure P2P services, and hybrid P2P services. For centralized P2P services, there is a central server in P2P services systems. However, the central server is not used as a central UDDI registry server; instead it is used for storing metadata of services to facilitate services discovery, which includes business names, services types, URLs, and so forth. In pure P2P services systems, services UDDI registry is distributed on every services node, so there is no need for services publication of the ordinary sense, and UDDI registry maintenance is also simplified because all services information is published and maintained locally. And in hybrid P2P services systems, the supernodes will be used for storing services metadata. It is useful for services discovery to cluster services nodes based on metadata, and then register the nodes in the same cluster under the same supernode.

BP-SERVICES: BESTPEER-BASED WEB SERVICES

As mentioned, the BP-Services project aims to develop an experimental P2P-based Web services platform as a test-bed for further P2P and Web services research.

BestPeer (Ng et al., 2002) is a generic P2P system with an architecture more pure P2P than hybrid P2P. The BestPeer system consists of two types of nodes: a large number of normal computers (i.e., peers), and a relatively fewer number of *Location-Independent Global names Lookup* (LIGLO) servers. Every peer in the system runs the BestPeer software, and will be able to communicate and share resources with any other peers. There are two types of data in each peer: private data and public (or sharable) data. For a certain peer, only its public data can be accessed by and shared with other

Figure 2. The internals of a BP-Services peer node



peers. Figure 1 shows the network topology of BestPeer. In the top layer are LIGLO servers, and in the bottom layer are normal peers.

The Architecture of BP-Services

In BP-Services, except for the LIGLO servers adhering to BestPeer, each peer node takes both the roles of a services provider and a services consumer, as well as a services registrar. That is to say, there is no central UDDI registry in BP-Services; all services and their definitions are distributed over the peer nodes. Figure 2 illustrates the internals of a BP-Services peer node. There are essentially seven components that are loosely integrated.

The first component, also the most important component, is the *Services Manager* that facilitates services discovery, services composition, and services deploying. Corresponding to its functionalities, the services manager consists of three sub-components: the *services discovery engine,* the *services composer,* and the *services deployer.* The services discovery engine is responsible for the publication and location of services. The services composer provides facilities for defining new composite services from existing services, and editing existing services (local). The services deployer facilitates the binding and invocation of requested services, as well as coordination of composite services.

The second component is the *Web Services Agent System,* or simply WSAgent. The WSAgent provides the environment for mobile agents to operate on. Each BP-Services node has a master agent that manages the services discovery and services description retrieval. In particular, it will clone and dispatch worker agents to neighboring nodes, receive results, and present to the user. It also monitors the statistics and manages the network reconfiguration policies.

The third component is a *Cache Manager,* which is used for caching the results of services discovery and retrieval. Furthermore, by collaboration among the cache managers, a P2P cache subsystem can be formed under the BP-Services framework so that all peers can share the caching results among themselves.

The fourth component is the *User Interface,* which consists of several interface modules, corresponding to services discovery and retrieval, services composition, and deploying.

The other three components are *Services Indexes*, *Local UDDI Registry,* and *Local Services Repository* respectively. The services repository keeps the services provided locally. Local UDDI registry holds the description (or publication) information of local services. And services indexes are simply inverted lists of services keywords extracted from the description information of local services, mainly business names and service types. Extracting and keeping services keywords speeds up services discovery.

Neighbor Nodes Finding in BP-Services

In BP-Services, given a participant node, its neighbor nodes are defined as those nodes that can provide as many services as possible similar to that in the given node. Here we use the information retrieval method to find neighbor nodes.

We can treat each node in BP-Services as a document, whose content is the services description information (UDDI registry) contained in that node. Thus we can cluster the nodes in BP-Services by using documents clustering methods (Baesa-Yates & Ribeiro-Neto, 1999). Roughly speaking, nodes in the same cluster may provide more similar services than those from different clusters. However, traditional documents clustering methods are based on a global data view, which is not realistic because it is not easy, if not impossible, to gather data of all nodes in a dynamic P2P network. In BP-Services, we adopt a simple local clustering strategy. We use a Boolean model to represent a peer services node, for the Boolean model is easier to evaluate than the vector space model (VSM), and it is difficult to set the document vector space without deterministic global data view in a P2P environment.

Given a peer services node *p*, there exists a set of keywords extracted from the services description document of *p*. We denote the set of keywords K_p, and treat it equal to the node *p* itself. For two services nodes *p* and *q*, suppose their keyword sets are K_p and K_q; the similarity of the two nodes are defined as follows. Here, |•| indicates the cardinality of a set.

$$sim(K_p, K_q) = \frac{|K_p \cap K_q|}{|K_p \cup K_q|}. \tag{1}$$

As in BestPeer, when a services node would like to become a participant of BP-Services, then it first registers with a LIGLO server, and the LIGLO server will issue the node with a global and unique identifier (i.e., BPID, BestPeerID), and meanwhile the LIGLO server will also send the node a list of peer nodes that have already registered in the network (i.e., the initial direct peers of the node). We term the links between these initial direct peers and the new participant node the *initial links* of the node.

After joining the network, the node (say *p*) can begin to find its neighbors by the following steps: (1) Through the ping-pong messages, it contacts the set of peers within a certain number (say *k*) of hops away from it. Let denote the set of peers as Peer(*p*, *k*)=$\{q_1,q_2,\ldots q_n\}$, and get these peers' keywords sets $\{K_i|i=1\sim n\}$. (2) Calculate the similarity of *p* and each peer in *Peer*(*p*, *k*)—that is, $\{sim(p, q_i)|\ i=1\sim n\}$. (3) Suppose *q* is the peer in *Peer*(*p*, *k*) that has largest similarity with *p*, then take *q* as *p*'s neighbor node, and connect *p* and *q* by a direct link, which is termed *neighbor link* of *p* and *q*.

Through the process of neighbor finding, the peers that share services tend to be connected together by neighbor links, and consequently form clusters of services peers. Considering the dynamism of the P2P system, the peers should update their neighbors regularly.

Services Discovery in BP-Services

Services discovery is the key process of P2P services. Because P2P services' UDDI registries are distributed on the peer nodes, it is inefficient to search the targeted services by traversing all peers one by one. Note that service discovery in P2P is different from P2P information retrieval. In service discovery, once a service that satisfies the requester's requirements is found, the discovery process can be stopped. It is not necessary to find a lot of similar services for a certain requester's specific service requirements.

In BP-Services, once a requester submits his or her service requirement, say a service query *Q*, the following process will be launched:

1. Extracting keywords from *Q*, the service search process is equal to carrying out keywords matching in information retrieval.
2. First, search at the local peer. The searching task is done by using the local services indexes as in traditional IR. If there are services matching the query, then go to (3); otherwise, go to (4).
3. Return the matched services' descriptions to the user, and the user browses the services descriptions to see whether there are services (s)he wants. If there is at least one service (s)he wants, then the process of service discovery is over; otherwise, go to (5).
4. Select randomly an *initial link* of the local peer, then clone a working agent and dispatch it with the service query to the peer at the other end of the selected initial link. At that remote peer, do the searching as at the local peer.
5. Clone a working agent and dispatch it with the service query to the local peer's neighbor. At the neighbor peer, do the searching as at the local peer.
6. At the remote peer, once there are services matching the query, then return the matching services' descriptions to the user, who decides whether the returned results contain the target service. If the target service is found, then the search task is over and the working agent would return the source peer or be destroyed at the remote peer. If no target service is found, the working agent has to continue the search target until the target service is found or the working agent's TTL is 0.

Note in the above process, when the working agent gets to a peer along a *neighbor link,* its TTL will not decrease; only walking along *initial link,* its TTL will decrease.

RELATED WORK

Recently, combining P2P and Web services is gaining importance both in industry and academia. From the industry, two ambitious projects were launched, Sun Microsystems' JXTA (Li, 2001) and Microsoft's .net, more recently Hailstorm. Both JXTA and Hailstorm are trying to provide a general, language/environment-independent P2P services environment by putting forward a set of protocols for communication among peers.

From research institutions, Hoschek (2002) proposed a unified peer-to-peer database framework for scalable service discovery; Schlosser Sinteck, Decker, and Nejdl (2002) put forward a scalable and ontology-based infrastructure for semantic Web services; Sheng, Benatallah, Dumas, and Mak (2002) developed a platform for rapid composition of Web services in peer-to-peer environment; and Abiteboul, Benjelloun, Manolescu, Milo, and Weber (2002) designed a kind of active XML document to integrate peer-to-peer data and Web services.

Unlike the projects above, BP-Services is based on the BestPeer platform. We use an information retrieval method for services discovery, which is quite different from other P2P services projects. BP-Services is easy to implement because, except for the ordinary Web services protocols, it does not need any additional and complex protocols.

CONCLUSION

To overcome the limitations of Web services systems caused by their centralized architecture, we explore the techniques of building Web services applications under a P2P environment. The ongoing project, BP-Services, is presented as a case study to demonstrate our approach. BP-Services is an experimental Web services platform developed on the propriety BestPeer infrastructure. Future work will focus on developing some concrete applications on BP-Services put on a campus network as a test-bed for future research on P2P and Web services. And semantic Web service will also be considered in BP-Services in the future.

ACKNOWLEDGMENTS

This work was supported by grants numbered 60573183 and 60373019 from the NSFC, grant No. 20045006071-16 from the Chenguang Program of Wuhan Municipality, grant No. WKL(04)0303 from the Open Researches Fund Program of LIESMARS, and the Shuguang Scholar Program of Shanghai Education Development Foundation.

REFERENCES

Abiteboul, S., Benjelloun, O., Manolescu, I., Milo, T., & Weber, R. (2002). Active XML: Peer-to-peer data and Web services integration. *Proceedings of the 28th International Conference on Very Large Databases* (pp. 1087-1090), Hong Kong, China.

Baesa-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (pp. 124-127). Boston: Addison-Wesley/ACM Press.

Christensen, E., Curbera, F., & Meredith, G. (2001). *Web Services Description Language (WSDL) 1.1.* W3C Note 15. Retrieved from http://www.w3.org/TR/wsdl

Gottschalk, K., Graham, S., Kreger, H., & Snell, J. (2002). Introduction to Web services architecture. *IBM Systems Journal, 41*(2), 170-177.

Hoschek, W. (2002). A unified peer-to-peer database framework and its application for scalable service discovery. *Proceedings of the 3rd International IEEE/ACM Workshop on Gird Computing,* Baltimore, MD (pp. 126-144).

Li, G. (2001). JXTA: A network programming environment. *IEEE Internet Computing,* (May-June), 88-95.

Ng, W. S., Ooi, B. C., & Tan, K.L. (2002). BestPeer: A self-configurable peer-to-peer system. *Proceedings of the 18th International Conference on Data Engineering* (pp. 272-272).

Ooi, B. C., Tan, K-L., Lu, H., & Zhou, A. (2002). P2P: Harnessing and riding on peers. *Proceedings of National Database Conference*, Zhengzhou, China, (pp. 1-5).

Schlosser, M., Sinteck, M., Decker, S., & Nejdl, W. (2002). A scalable and ontology-based infrastructure for semantic Web services. *Proceedings of the 2nd International Workshop on Agents and Peer-to-Peer Computing*, Linköping, Sweden, (pp. 104-111).

Sheng, Q., Benatallah, B., Dumas, M., & Mak, E. (2002). SELF-SERV: A platform for rapid composition of Web services in a peer-to-peer environment. *Proceedings of the 28th International Conference on Very Large Databases*, Hong Kong, China, (pp. 1051-1054).

SOAP. (2000). *Simple Object Access Protocol (SOAP) 1.1.* W3C Note 8. Retrieved from http://www.w3.org/TR/soap

Yang, B., & Garcia-Molina, H. (2001). Comparing hybrid peer-to-peer systems. *Proceedings of the 27th International Conference on Very Large Databases*, Roma, Italy, (pp. 561-570).

Web Services Conceptual Architecture. (n.d.). Retrieved from *http:*//www.ibm.com/software/solutions/webservices/documentation.html

KEY TERMS

Centralized P2P: In a centralized P2P network, there is a central server responsible for maintaining indexes on the metadata for all the peers in the network.

Hybrid P2P: A kind of tradeoff between centralized P2P and pure P2P, which is structured hierarchically with a supernode layer and a normal peers layer.

Peer-to-Peer (P2P): A class of systems and applications that employ distributed resources to perform a critical function in a decentralized way. A P2P distributed system typically consists of a large number of nodes that can share resources, information, and services, taking the roles of both consumer and provider, and may join or depart the network at any time, resulting in a truly dynamic and ad-hoc environment.

Pure P2P: A P2P system with fully autonomous peers—that is, all nodes are equal, no functionality is centralized, and the communication between peers is also symmetric.

Service Discovery: An operation of finding Web services. After Web services are created and published in Web services registries such as UDDI, the service users or consumers need to search Web services manually or automatically. The implementation of UDDI servers should provide simple search APIs or Web-based GUI to help find Web services.

Universal Description, Discovery and Integration (UDDI): The protocol for Web service publishing. It should enable applications to look up Web services information in order to determine whether to use them.

Web Service: Can be seen as an interface that describes a collection of operations that are network accessible through standardized XML messaging. Software applications written in various programming languages and running on various platforms can use Web services to exchange data over computer networks due to the interoperability of using open standards.

Web Services Description Language (WSDL): An XML language for describing Web services.

eXtensible Markup Language (XML): A meta-language written in SGML that allows one to design a markup language, used to allow for the easy interchange of documents on l is misd Wide Web.

Business and Technology Issues in Wireless Networking

David Wright
University of Ottawa, Canada

INTRODUCTION

A major development in the enabling technologies for mobile computing and commerce is the evolution of wireless communications standards from the IEEE 802 series on local and metropolitan area networks. The rapid market growth and successful applications of 802.11, WiFi, is likely to be followed by similar commercial profitability of the emerging standards, 802.16e, WiMAX, and 802.20, WiMobile, both for network operators and users. This article describes the capabilities of these three standards and provides a comparative evaluation of features that impact their applicability to mobile computing and commerce. In particular, comparisons include the range, data rate in Mbps and ground speed in Km/h plus the availability of quality of service for voice and multimedia applications.

802.11 WiFi

WiFi (IEEE, 1999a, 1999b, 1999c, 2003) was originally designed as a wireless equivalent of the wired local area network standard IEEE802.3, Ethernet. In fact there are many differences between the two technologies, but the packet formats are sufficiently similar that WiFi packets can easily be converted to and from Ethernet packets. Access points can therefore be connected using Ethernet and can communicate with end stations using WiFi.

WiFi can transport both real-time communications such as voice and video plus non-real time communications such as Web browsing, by providing quality of service, QoS, using 802.11e (IEEE, 2005). There are 2 QoS options. One provides four priority levels allowing real-time traffic to be transmitted ahead of non-real-time traffic, but with no guarantee as to the exact delay experienced by the real-time traffic. The other allows the user to request a specific amount of delay, for example, 10 msecs., which may then be guaranteed by the access point. This is suited to delay sensitive applications such as telephony and audio/video streaming.

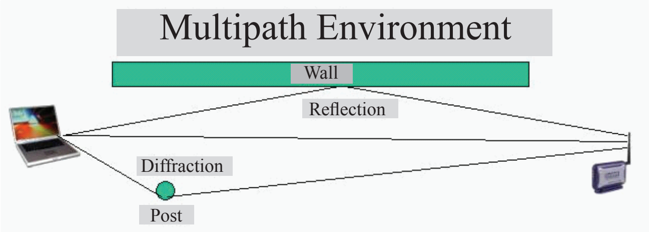
WiFi has a limited range of up to 100 metres, depending on the number of walls and other obstacles that could absorb or reflect the signal. It therefore requires only low powered transmitters, and hence meets the requirements of operating in unlicensed radio spectrum at 2.4 and 5 GHz in North America and other unlicensed bands as available in other countries.

WiFi is deployed in residences, enterprises and public areas such as airports and restaurants, which contain many obstacles such as furniture and walls, so that a direct line of sight between end-station and access point is not always possible, and certainly cannot be guaranteed when end stations are mobile. For this reason the technology is designed so that the receiver can accept multipath signals that have been reflected and/or diffracted between transmitter and receiver as shown in Figure 1(a). WiFi uses two technologies that operate well in this multipath environment: DSSS, Direct Sequence Spread Spectrum, which is used in 802.11b, and OFDM, Orthogonal Frequency Division Multiplexing, which is used in 802.11a and g (Gast, 2002). A key distinguishing factor between these alternatives, which is important to users, is spectral efficiency, that is, the data rate that can be achieved given the limited amount of wireless spectrum available in the unlicensed bands. DSSS as implemented in 802.11b uses 22 MHz wireless channels and achieves 11 Mbps, that is, a spectral efficiency of 11/22 = 0.5. OFDM achieves a higher spectral efficiency and is therefore making more effective use of the available wireless spectrum. 802.11g has 22 MHz channels and delivers 54 Mbps, for a spectral efficiency of 54/22 = 2.5 and 802.11a delivers 54 Mbps in 20 MHz channels, with a spectral efficiency of 54/20 = 2.7. A recent development in WiFi is 802.11n (IEEE, 2006a), which uses OFDM in combination with MultiInput, MultiOutput, MIMO, antennas as shown in Figure 1(b). MIMO allows the spectral efficiency to be increased further by exploiting the multipath environment to send several streams of data between the multiple antennas at the transmitter and receiver. At the time of writing the details of 802.11n are not finalized, but a 4x4 MIMO system (with 4 transmit and 4 receive antennas) will probably generate about 500 Mbps in a 40 MHz channel, that is, a spectral efficiency of 500/40 = 12.5. 802.11n is suited to streaming high definition video and can also support a large number of users per access point.

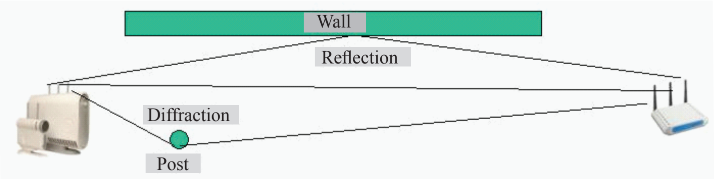
The data rates in WiFi are shared among all users of a channel, however some users can obtain higher data rates than others. Network operators may choose to police the data rate of individual users and possibly charge more for higher rates, or they may let users compete so that their data rates vary dynamically according to their needs and the priority levels

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. (a) Receiver recovers a single signal from multiple incoming signals; (b) MIMO receiver recovers multiple signals using multiple antennas



(a)



(b)

of their traffic. This provides considerable flexibility allowing many users to spend much of their time with low data rate applications such as VoIP, e-mail and Web browsing, with occasional high data rate bursts for audio/video downloads and data-intensive mesh computing applications.

Many deployments of WiFi use multiple access points to achieve greater coverage than the range of a single access point. When the coverage of multiple access points overlaps they should use different radio channels so as not to interfere with each other, as shown in Figure 2. For instance, in the North American 2.4 GHz band there is 79 MHz of spectrum available and the channels of 802.11b and g are 22 MHz wide. It is therefore possible to fit 3 non-overlapping channels into the available 79 MHz, which are known as channels 1, 6 and 11. Other intermediate channels are possible, but overlap with channels 1, 6 and 11. In Figure 2, the top three access points are shown connected by Ethernet implying that they are under the control of a single network operator, such as an airport. As an end-station moves among these access points the connection is handed off from one access point to another using 802.11r (IEEE, 2006b), while maintaining an existing TCP/IP session. Movement can be up to automobile speeds using 802.11p (IEEE, 2006c). Standard technology, 802.21 (IEEE, 2006d), is also available to handoff a TCP/IP session when a mobile end-station moves from an access point of one network operator to that of another, and this requires a business agreement between the two operators.

802.11 networks can therefore span extensive areas by interconnecting multiple access points, and city-wide WiFi networks are available in, for example, Philadelphia in the U.S., Adelaide in Australia, Fredericton in Canada and Pune in India. The broad coverage possible in this way greatly expands the usefulness of WiFi for mobile computing and electronic commerce. Enterprise users can set up secure virtual private networks from laptops to databases and maintain those connections while moving from desk to conference room to taxi to airport. A VoIP call over WiFi can start in a restaurant, continue in a taxi and after arriving at a residence.

The features of WiFi, IEEE 802.11, that are of particular importance for mobile computing and commerce are:

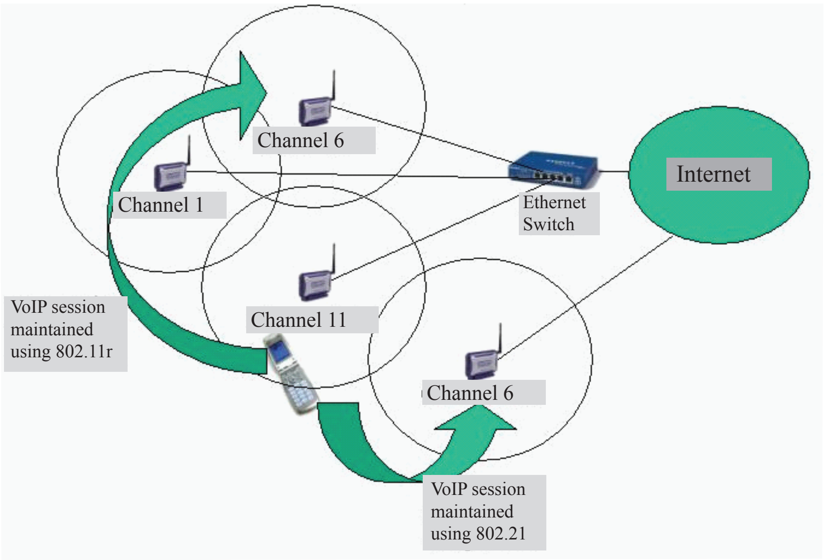
- Broad coverage achieved by handing off calls between access points, using 802.11r and 802.21, in cities where there are sufficient access points.
- Multimedia capability achieved by QoS, 802.11e.
- Flexibility in data rates achieved by allowing the total data rate of an access point to be shared in dynamically changing proportions among all users.
- Low cost achieved by using unlicensed spectrum, low power transmitters and mass produced equipment.

The downside to WiFi, IEEE 802.11, is limited coverage in cities that do not have extensive access point deployment.

802.16E WIMAX

802.16E (IEEE, 2006e) has a greater range than 802.11, typically 2-4 km and operates between base stations and subscriber stations. The initial IEEE standard 802.16 is for fixed applications, which compete with DSL and cable

Figure 2. WiFi handoff among access points



modems. Mobile applications including handoff capability among base stations, which we deal with here, are provided by 802.16E, and are based on similar but incompatible technology.

In 802.16E, WiMAX, mobility is limited to automobile speeds, up to about 100 Km/h so that it has limited use in high speed trains and aircraft. WiMAX uses the terminology "subscriber" stations, implying that customers are paying for a public service. Since the geographic range extends well into public areas, this is certainly one application. Another mobile application is a private campus network in which a central base station serves a business park or university campus. Initial deployment of WiMAX uses licensed spectrum, although low power applications in unlicensed spectrum are also specified in the standard.

WiMAX has sophisticated QoS capabilities, which allow customers to reserve capacity on the network including a reserved data rate plus quality of service. The data rate is specified by a minimum reserved traffic rate, MRTR, on which quality of service is guaranteed (Figure 3). The customer is allowed to send at a higher rate, up to a maximum sustainable traffic rate, MSTR, without necessarily receiving QoS, and above that rate, traffic will be policed by the network operator, that is, it may be discarded. The QoS parameters that can be specified by the customer are latency and jitter, plus a priority level, which is used by the base station to distinguish among service flows that have the same latency and jitter requirements. The combination of latency and jitter can be used to distinguish among service flows, and further detail on the performance of WiMAX is given by Ghosh et al. (2005).

Combinations of QoS parameters and data rates make WiMAX highly suited to mobile computing and commerce. Each subscriber can set up multiple service flows, for example, for Web browsing during a multimedia conference, and use data rates that are quite different from those of other customers. The service provider can charge based on a combination of data rate and QoS.

WiMAX is based on OFDM, thus achieving a high spectral efficiency. There are a number of options within 802.16E for the channel widths and modulation techniques, resulting in a corresponding range of data rates and spectral efficiencies. It is important to recognize that the spectral efficiency depends on the distance between the base station and the subscriber station (Figure 4). As the signal degrades with distance it is not possible to encode so many bps within each Hz and 802.16E assigns encodings that take this into account. Closer to the base station the data rate is therefore higher. The exact distance depends on the operating environment

Figure 3. WiMAX traffic rate guarantees

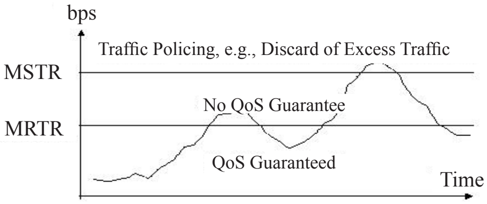
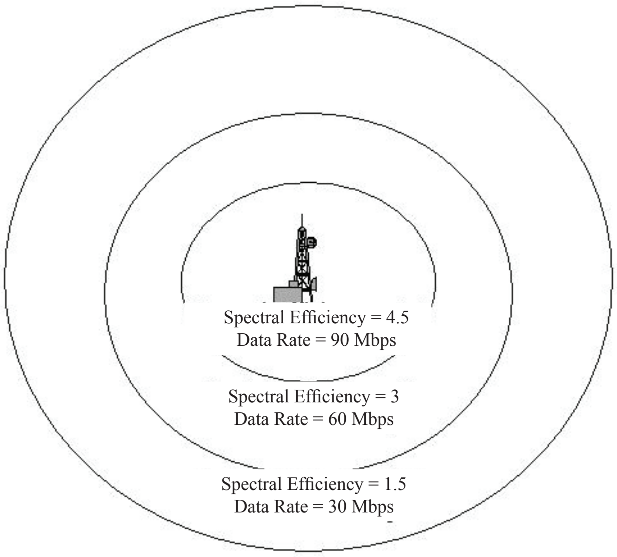


Figure 4. Spectral efficiency and maximum data rates for WiMAX



since 802.16E uses multipath signals involving reflections and diffractions. The data rates shown in Figure 4 are the maximum achievable with the highest channel bandwidth allowed according to the standard—20 MHz—and can vary not only with distance but also according to how much forward error correction is used.

The features of 802.16E that are of particular importance for mobile computing and commerce are:

- Good range, enabling city-wide coverage with a reasonable number of base stations.
- Multimedia capability achieved by QoS, and guaranteed data rates.
- Flexibility in data rates achieved by allowing the total data rate of a base station to be shared in dynamically changing proportions among all users.

The downside to 802.16E is the cost of licensed spectrum.

802.20 WIMOBILE

At the time of writing, (1Q06), the specification of 802.20, (IEEE, 2006, f), is under development, so that less detail is available than for 802.11 and 802.16e. The key features of 802.20 are:

- It operates in licensed spectrum below 3.5 GHz.
- It is designed from the start for an all-IP environment and interfaces to IP DiffServ QoS service classes, (Grossman, 2002) which provide for prioritization of users' traffic.
- It interfaces to "Mobile IP" (Montenegro, 2001) as part of its mobility capability. Mobility includes not just automobile speed, but also high speed trains at up to 250 Km/h.
- It uses OFDM with MIMO antennas to achieve a very high spectral efficiency, so that large numbers of users can share access to a single base station.

COMPARATIVE EVALUATION

Mobile computing and commerce involves communicating from mobile devices for a variety of purposes including: data transfer for processing intensive applications and for Web browsing; voice and multimedia calls between human users;

Table 1. Comparative evaluation of technologies for mobile computing and commerce

	802.11, WiFi	802.16e, WiMAX	802.20, WiMobile
Range	100 metres	2-4 Km	2-4 Km
Coverage	Hot spots. Some city-wide deployments.	Designed for city-wide deployment	Designed for national deployment
Data Rate	11, 54, 500 Mbps flexibly shared among all users	Up to 90 Mbps flexibly shared among all users	> 1 Mbps per user
QoS	(a) Prioritization mechanism (b) data rate and QoS guarantees	Data rate and QoS guarantees	Data rate guarantees and QoS prioritization
Mobility Speed	100 Km/h	100 Km/h	250 Km/h
Cost	Very low unit cost access points. End-station interfaces built into phones, laptops, PDAs. Large number of access points required. Unlicensed spectrum.	Medium unit cost access points. End-station interfaces built into phones, laptops, PDAs. Licensed or unlicensed spectrum.	Medium unit cost access points. End-station interfaces built into phones, laptops, PDAs. Licensed spectrum.

downloading audio, video and multimedia from a server, (a) streaming for real-time playout to human users and (b) file transfer for subsequent access on the mobile device. Each of these requires appropriate data rate and quality of service. Cost is also an important factor, since subscription may be required to a public network operator or an enterprise may need to build its own wireless network. Employees using mobile computing devices within a building require mobility only at pedestrian speeds. In public areas such as city streets, automobile speeds are required and between cities high speed trains may be used. The type of mobile computing application determines which speed is appropriate. Table 1 provides a comparison among the three technologies described in this paper.

CONCLUSION

Mobile computing and commerce users have a wide range of emerging wireless communication technologies available: WiFi, WiMAX and WiMobile. Each of them offers high data rates and spectral efficiencies, and will therefore likely be available at low cost. They are the major enabling telecommunication technologies for mobile computing and are likely to be deployed in public areas and private campuses for in-building and outdoor use. WiFi is already extensively deployed and WiMAX is being deployed in Korea in 2006 and can be expected in many other countries in 2007. The WiMobile standard has not yet been specified (as of the time of writing 1Q06) and commercial equipment can be expected after WiMAX.

REFERENCES

Gast, M. (2002). *802.11 wireless networks: The definitive guide***.** O'Reilly.

Ghosh, A., Wolter, D. R., Andrews, J. G., & Chen, R. (2005). Broadband wireless access with WiMax/8O2.16: Current performance benchmarks and future potential. *IEEE Communications Magazine*, *43*(2), 129-136.

Grossman, D. (2002). *New terminology and clarifications for Diffserv*. RFC3260. Internet Engineering Task Force.

IEEE. (1999a). *802.11 wireless LAN: Medium access control (MAC) and physical layer (PHY) specifications*. New York: IEEE Publications.

IEEE. (1999b). *802.11a high-speed physical layer in the 5 GHz band*. New York: IEEE Publications.

IEEE. (1999c). *802.11b higher-speed physical layer (PHY) extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2003). *802.11g further higher-speed physical layer extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2005). *802.11e wireless LAN: Quality of service enhancements*. New York: IEEE Publications.

IEEE. (2006a). *802.11n wireless LAN: Enhancements for higher throughput* (In progress). Retrieved March 2006, from http://standards.ieee.org/board/nes/projects/802-11n.pdf.

IEEE. (2006 b). *802.11r wireless LAN: Fast BSS transition* (In progress). Retrieved March 2006, http://standards.ieee.org/board/nes/projects/802-11n.pdf.

IEEE. (2006c). *802.11p wireless LAN: Wireless access in vehicular environments*. (In progress). Retrieved March 2006, from http://standards.ieee.org/board/nes/projects/802-11p.pdf.

IEEE. (2006d). *802.21 media independent handover services*. (In progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/21/.

IEEE. (2006e). *802.16E-2005 air interface for fixed and mobile broadband wireless access systems: Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands*. New York: IEEE Publications.

IEEE. (2006f). 802.20 mobile broadband wireless access systems. (In progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/20/.

Montenegro, G. (2001) *Reverse tunneling for mobile IP*. RFC3024. Internet Engineering Task Force.

KEY TERMS

Direct Sequence Spread Spectrum (DSS): A transmission technique in which data bits are multiplied by a higher frequency code sequence, so that the data are spread over a wide range of frequencies. If some of these frequencies fade, the data can be recovered from the data on the other frequencies together with a forward error correction code.

Mobile IP: An Internet standard that allows a mobile user to move from one point of attachment to the network to another while maintaining an existing TCP/IP session. Incoming packet to the user are forwarded to the new point of attachment.

Multipath: A radio environment in which signals between transmitter and receiver take several different spatial paths due to reflections and diffractions.

Orthogonal Frequency Division Multiplexing (OFDM): A transmission technique in which data bits are transmitted on different frequencies. The data transmitted on one frequency can be distinguished from those on other frequencies since each frequency is orthogonal to the others.

Quality of Service (QoS): Features related to a communication, such as delay, variability of delay, bit error rate and packet loss rate. Additional parameters may also be included, for example, peak data rate, average data rate, percentage of time that the service is available, mean time to repair faults and how the customer is compensated if QoS guarantees are not met by a service provider.

WiFi: A commercial implementation of the IEEE 802.11 standard in which the equipment has been certified by the WiFi Alliance, an industry consortium.

WiMAX: A commercial implementation of the IEEE 802.16 standard in which the equipment has been certified by the WiMAX Forum, an industry consortium.

WiMobile: Another name for the IEEE 802.20 standard which is in course of development at the time of writing (1Q06).

Business Strategies for Mobile Marketing

Indranil Bose
University of Hong Kong, Hong Kong

Chen Xi
University of Hong Kong, Hong Kong

INTRODUCTION

With the appearance of advanced and mature wireless and mobile technologies, more and more people are embracing mobile "things" as part of their everyday lives. New business opportunities are emerging with the birth of a new type of commerce known as mobile commerce or m-commerce. M-commerce is an extension to electronic commerce (e-commerce) with new capabilities. As a result, marketing activities in m-commerce are different from traditional commerce and e-commerce. This chapter will discuss marketing strategies for m-commerce. First we will give some background knowledge about m-commerce. Then we will discuss the pull, push, and viral models in m-marketing. The third part will be the discussion about the future developments in mobile marketing. The last part will provide a summary of this article.

BACKGROUND

Popularity of Mobile Services

From the research done by Gartner Dataquest (BusinessWeek, 2005), there will be more than 1.4 billion mobile service subscribers in the Asia-Pacific region by 2009. Research analysts of Gartner Dataquest also estimated that China will have over 500,000 subscribers, and more than 39% of the people will use mobile phones at that time. In India, the penetration rate of mobile phones is expected to increase from 7% in 2005 to 28% in 2008. The Yankee Group has also reported a growing trend of mobile service revenues from 2003 to 2009. Although the revenue generated by traditional text-based messaging service will not change much, revenue from multimedia messaging services will rise to a great extent. Other applications of mobile services, such as m-commerce-based services and mobile enterprise services, will continue to flourish. One thing that is very important in driving Asia-Pacific mobile service revenue is mobile entertainment services. Revenue from mobile entertainment services will make up almost half of the total revenues from all kinds of mobile data services from now on. Not only in the region of Asia-Pacific, but mobile services will increase in popularity in other parts of the world as well. In the United States, it is expected that the market for m-commerce will reach US$25 billion in 2006.

The Development of Mobile Technologies

Two terms are frequently used when people talk about mobile information transmission techniques: the second-generation (2G) and the third-generation (3G) wireless systems. These two terms actually refer to two generations of mobile telecommunication systems. Three basic 2G technologies are time division multiple access (TDMA), global system for mobile (GSM), and code division multiple access (CDMA). Among these three, GSM is the most widely accepted technology. There is also the two-and-a-half generation (2.5G) technology of mobile telecommunication, such as general packet radio service (GPRS). 2.5G is considered to be a transitional generation of technology between 2G and 3G. They have not replaced 2G systems. They are mostly used to provide additional value-added services to 2G systems. The future of mobile telecommunication network is believed to be 3G. Some standards in 3G include W-CDMA, TD-SCDMA, CDMA 2000 EV-DO, and CDMA EV-DV. The advancement in mobile telecommunication technology will bring in higher speed of data transmission.. The speed of GSM was only 9.6 kilobits per second (kbps), while the speed of GPRS can reach from 56 to114 kbps. It is believed that the speed of 3G will be as fast as 2 Megabits per second (mbps). The acceptance of 3G in this world began in Japan. NTT DoCoMo introduced its 3G services in 2001. Korea soon followed the example of Japan. In 2003, the Hutchison Group launched 3G commercially in Italy and the UK, and branded its services as '3'. '3' was later introduced in Hong Kong, China in 2004. Mainland China is also planning to implement 3G systems. Some prototypes or experimental networks have been set up in the Guangdong province. It is expected that 3G networks will be put into commercial use in 2007 using the TD-SCDMA standard that has been indigenously developed in China. Mobile information transmission can also be done using other technical solutions such as wireless local area network (WLAN) and Bluetooth. The interested reader may refer to Holma and Toskala (2002) for a fuller

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

description of 3G systems, and to Halonen, Romero, and Melero (2003) for details of 2G and 2.5G systems.

The most popular mobile devices currently in use include mobile phones, wireless-enabled personal digital assistants (PDAs), and wireless-enabled laptops (Tarasewich, Nickerson, & Warkentin, 2002). Smartphones are also gaining favor from customers. Mobile phones are the most pervasive mobile devices. Basically, mobile phones can make phone calls, and can send and receive short text messages. More advanced mobile phones have color screens so that they can send or receive multimedia messages, or have integrated GPRS modules so that they can connect to the Internet for data transmission. PDAs are pocket-size or palm-size devices which do limited personal data processing such as recording of telephone numbers, appointments, and notes on the go. Wireless-enabled PDAs have integrated Wi-Fi (wireless fidelity)—which is the connection standard for W-LAN or Bluetooth—which helps them access the Internet. Some PDAs can be extended with GPRS or GSM modules so that they can work as a mobile phone. PDAs nowadays usually have larger screens than that of mobile phones and with higher resolution. They are often equipped with powerful CPUs and large storage components so that they can handle multimedia tasks easily. Smartphones are the combination of mobile phones and PDAs. Smartphones have more complete phoning function than PDAs, while PDAs have more powerful data processing abilities. However, the boundary between smartphones and PDAs are actually becoming more and more fuzzy.

The Need for Mobile Marketing

The rapid penetration rate of mobile devices, the huge amounts of investment from industries, and the advancement of mobile technologies, all make it feasible to do marketing via mobile devices. Mobile commerce refers to a category of business applications that derive their profit from business opportunities created by mobile technologies. Mobile marketing, as a branch of m-commerce (Choon, Hyung, & Kim, 2004; Varshney & Vetter, 2002), refers to any marketing activities conducted via mobile technologies. Usually m-commerce is regarded as a subset of e-commerce (Coursaris & Hassanein, 2002; Kwon & Sadeh, 2004). That is true, but due to the characteristics of mobile technologies, mobile marketing is different from other e-commerce activities. The first difference is caused by mobile technologies' ability to reach people anywhere and anytime; therefore mobile marketing can take the advantage of contextual information (Zhang, 2003). Dey and Abowd (2001) defined context as "any information that characterizes a situation related to the interaction between users, applications, and the surrounding environment." Time, location, and network conditions are three of the key elements of context. The second difference is caused by the characteristics of mobile devices. Mobile devices have limited display abilities. The screens are usually small, and some of the devices cannot display color pictures or animations. On the other hand, mobile devices have various kinds of screen shapes, sizes, and resolutions. Thus, delivering appropriate content to specific devices is very important. Mobile devices also have limited input abilities, and this makes it difficult for customers to respond. Mobile marketing shares something in common with e-commerce activities. An important aspect of e-commerce is to deliver personalized products/services to customers. Mobile marketing inherits this feature. Mobile marketing also inherits some of the problems from e-commerce, especially the problem of spamming. Personalization in mobile marketing is to conduct marketing campaigns which can meet the customer's needs by providing authorized, timely, location-sensitive, and device-adaptive advertising and promotion information (Scharl, Dickinger, & Murphy, 2005).

MOBILE MARKETING

Benefits of Mobile Marketing

There are two main approaches to advertise and promote products in industry—mass marketing and direct marketing. The former uses mass media to broadcast product-related information to customers without discrimination, whereas the latter is quite different in this regard. Mobile marketing takes a direct marketing approach. Using mobile marketing, marketers can reach customers directly and immediately. Similarly, customers can also respond to marketers rapidly. This benefit makes the interaction between marketers and customers easy and frequent. Compared to direct marketing using mail or catalogs, mobile marketing is comparatively cost effective and quick. Compared to telephone direct marketing, mobile marketing can be less interruptive. Compared to e-mail direct marketing, mobile marketing can reach people anytime and anywhere, and does not require customers to sit in front of a computer. Therefore, to some extent, mobile marketing can be a replacement for other types of marketing channels such as mail, telephone, or e-mail. Advertisement or promotion information sent via the Internet can be sent via a mobile device. Mobile marketing can enhance marketing by adding new abilities like time-sensitive and location-sensitive information. On the other hand, mobile commerce can generate new customers' data, like mobile telecommunication usage data and mobile Internet surfing data. Mobile marketing is the first choice for conducting marketing activities for m-commerce applications. However, due to limited size of screens of mobile devices, only brief information can be provided in mobile marketing solicitations, while e-mail or mail marketing can provide very detailed information. On the other hand telephone marketing requires the good communication skill of telesales. Once telemarketers have acquired this skill, the interaction between marketers and customers is quicker and more effective. It is not clear if mobile marketing is as effective or as popular as mass marketing agents like television and newspaper, but it can be said that it is indeed a powerful medium that is likely to gain in popularity in the future.

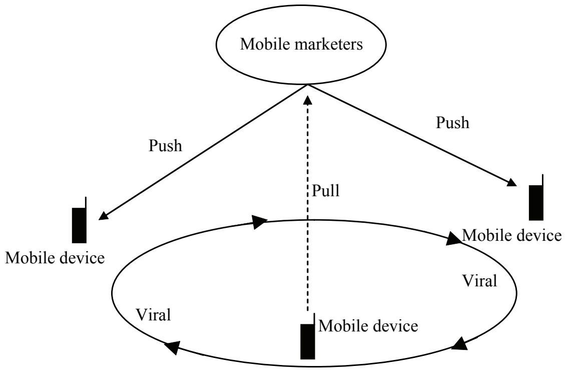
Models for Mobile Marketing

Mobile marketing usually follows one of the three kinds of models—push, pull (Haig, 2002; Zhang, 2003), and viral (Ahonen, 2002; Ahonen, Kasper, & Melkko, 2004; Haig, 2002), as illustrated in Figure 1.

Push Model

The push model sends marketing information to customers without the request of the customer. If the push model is used, besides knowing the targeted customers' interests, understanding the context of customers at the time the marketing activities are to be carried out is very critical. The timing in sending mobile information should also be appropriate. The content delivered to customers should also be displayable on their mobile devices. Permissions from customers are necessary before any solicitations can be sent. In the push model, marketers like to make it easier for customers to respond because of the poor input ability of mobile devices. G2000, a Hong Kong clothing chain, launched a mobile marketing campaign in November 2004. Mobile coupons in the format of SMS were sent to the mobile phones of selected customers. Customers could then use these coupons stored in their mobile phones when purchasing items in designated G2000 stores in order to get discounts. The campaign was considered to be a success because a number of customers responded to this program and used mobile coupons at G2000 stores.

Figure 1. Push, pull and viral models for mobile marketing



Pull Model

In the pull model, the marketer waits for the customer to send a request for a solicitation. The marketer prepares the marketing information in a format that is displayable in all possible mobile devices and scalable for various connection speeds. The significance of the pull model is that the information from customers is very useful for understanding customers' preferences, such as the preferred marketing time and interests. Mobile marketing following a pull model can be conducted in many ways. One possible approach is to let the customers select and download coupons to their mobile devices. Mobile service providers can build a Web site using a mobile Internet protocol such as wireless application protocol (WAP), and place various text-based coupons on this Web site. Customers can use their GPRS-enabled phone to browse the Web site and download coupons they like in the format of SMS. Each coupon will have a unique identity number. When the coupon is redeemed, related information, such as the phone number of the customer who downloaded the coupon and the time it was redeemed, is recorded. This kind of information is later used to analyze the behaviors of customers and build a profile for the customer. China Mobile, a mobile telecommunication operator in China, had established such a Web site for customers in Xiamen (a city located in southeast China) in 2005. The customers of China Mobile could download coupons displayed on this Web site onto their mobile phones using text messaging.

Viral Model

The phrase viral marketing was created by Steve Jurvetson in 1997 to describe the burgeoning use of Hotmail (Jurvetson & Draper, 1997). The principle of the viral model is based on the fact that customers forward information about products/services to other customers. The viral model enlarges the effect of other marketing activities while it costs the marketers very little in monetary terms. The viral model enables customer-to-customer communication. Like the pull model, the format of information that is delivered by viral model should be displayable in different devices and scalable for different connection speeds. Actually mobile marketing has the ability to be viral inherently because it is quite easy for people to forward mobile advertising or promotion information to their friends. However, viral marketing information has to be interesting and attractive enough to make the customers willing to forward it to other people. For example, "reply to this message in order to win $5000" may be a very attractive viral marketing message. Usually, viral marketing begins with push marketing activities to customers. According to Linner (2003), when the movie "2 Fast 2 Furious" was running in movie theatres, marketers tried to create a viral promotion using a mobile marketing strategy. Fans were asked to send SMS to enter a certain film-related competition. Besides inviting fans on every major phone network through advertisements on television, newspapers, and also through posters, a special code was designed and a low fee was offered to customers in order to encourage them to forward promotion information to their friends. Exciting gifts were offered as prizes in this competition (such as a replica of the vehicle, the EvO VII, that was used in the movie) to spur the enthusiasm of customers.

These three models of direct marketing can be complimentary to each other. Push-based mobile marketing can be used to stimulate pull-based marketing activities. For example, book marketers can send a short introduction to customers via SMS with a remark at the end saying "for more details, please reply to XXXXX." Once a customer responds to this by replying using SMS, more promotion or advertising information on this book can be sent to him. All three models—push, pull, and viral—can even be integrated together in a mobile marketing campaign. An example of an integrated mobile marketing approach was adopted by Fox Txt Club for the movie "phone booth" (Linner, 2003). At the beginning of the marketing campaign, members of the Fox Txt Club were sent invitations via SMS to a preview. The aim of this was to pull customers to the campaign. A competition that invited people to send SMS about questions pertaining to various details in the film was set up. The forwarding of SMS about the movie and the competition among club members and their friends, together with other media such as entertainment and event listings magazines and city-center posters, made the marketing campaign viral. The details of those who responded were recorded by Fox Txt Club, and this helped in building the database of customers for future release promotions. This could be used for push marketing for another movie in the future.

Strategies for Mobile Marketing

The most fundamental task for marketing activities following the push model is to send advertising or promotion information about products or services that the targeted customers once bought. This is the most direct and easy way to decide what is to be offered to customers in a solicitation. However, just marketing products already existing in customers' transaction records is not enough for marketers. It is necessary for marketers to explore the needs of customers. Two of the most commonly used marketing strategies are cross-selling and up-selling. Cross-selling is the practice of suggesting similar products or services to a customer who is considering buying something, such as showing a list of ring tones on a mobile Internet Web page that are similar to the one a customer has downloaded. Up-selling is the practice of suggesting higher priced, better versions of products or services to a customer who is considering a purchase, such as a mobile phone plan with higher fees and additional features. Two approaches can be used to find opportunities for up-selling or cross-selling. One is to find products or services that are similar to the ones a customer has bought. The other is to find people who have characteristics that are similar to a targeted customer. Products or services those people have bought and the targeted customer has not can be recommended to the target customers.

Pull-based marketing is relatively passive compared to push-based marketing. Usually in pull marketing, customers are responsible for searching for useful advertising or promotion information. The marketers' responsibility is to help customers find what they want more efficiently. Therefore, knowing what customers may request is very important in pull marketing. Instead of sending related information to customers like push marketing, marketers doing pull marketing can make information about products or services available on their mobile Internet Web site or ordinary Internet Web site. In viral marketing, marketers stand in a more passive position than even in pull marketing. However, for both pull and push marketing, some push activities should be carried out to start the marketing.

Whatever model one may use when carrying out mobile marketing activities, one issue must always be kept in mind and that is the necessity of obtaining explicit permission from customers (Bayne, 2002). Mobile technology makes connections so direct that it can interfere with customers' privacy very easily. Therefore, sending advertising or promotion information to people will cause trouble if permissions are not sought before solicitations or customers' wishes about not receiving a solicitation are not respected.

Understanding Customers in Mobile Marketing

All of the three models require good understanding of customers' needs. Marketing information that is not well designed will be regarded as spam by customers. Once a customer identifies some information from a company as spam, he or she will pay very little attention to or simply discard any information from that company. If a customer cannot find useful information on the Web site a company provides, it may be ok for the first time, a pity for the second time, but for the third time it will mean business lost forever. If information sent to customers is not interesting, customers may not want to forward them to their friends. All these situations may lead to failure of a marketing campaign. To avoid these situations, marketers need to understand customers well enough in order to send personalized marketing information. Customer profiling is a necessary approach to understand customers better. Customer profiling aims to find factors that can characterize customers. These factors are found by comparing customers to each other in order to discover similarities and differences among customers. Customer profiling encompasses two tasks—customer clustering and customer behavior pattern recognition. Customer clustering aims to classify customers into different groups. Customers within the same group are said to be more similar to each other than to customers in different groups. Marketers cluster customers using various data. Traditionally, customers are clustered according to their geographic locations, demographic characteristics, and the industries they are working for. They can also be clustered based on information about their purchasing history, such as what they bought, when they bought, and how much they spent. With the appearance of mobile services and m-commerce, usage data of new customer data services can also be used for clustering. For example, messaging services that customers subscribed to, GPRS surfing and download records, the type of mobile devices the customers use, and monthly mobile phone usage including use of IDD and roaming can yield many interesting information about the customers.

Aside from these hard facts, marketers may also want to infer some soft knowledge about customers' behaviors as well. To recognize customer behavior the marketers must discover relationships between hard facts. For example, customers that download ring tones of game music may download games-related screensavers later on. Since mobile technologies can enable context-sensitive marketing activities, marketers should gather knowledge about customers' location preferences and time preferences. For example, when does a customer usually go shopping and which place does he/she visit on the shopping trips? Marketers can find this kind of soft knowledge from various mobile network usage data. Again, collecting information on location and time requires permission from customers. Based on customer profiling, more sophisticated personalized advertising or promotion information can be sent to customers.

FUTURE TRENDS

Mobile technologies will advance further in the future. New technologies will enable new kinds of marketing activities. For example, the implementation of fourth-generation (4G) wireless systems will make the bandwidth much larger than that in current networks. On the other hand, the mobile device will have larger screens with higher resolutions. These two factors together will make interactive audio and even interactive video marketing possible. Generally speaking, the limitation of current mobile technologies will be weakened or removed in the future. As a result, more emphasis may be put on time- and location-related marketing, as well as on better understanding customers' interests. The principle is not only to know what customers want, but also to know when and where they may have a certain kind of need. Data mining techniques can be used in the future to find customer behavior patterns with time and location factors. Data mining techniques have been used widely in direct marketing for targeting customers (Ling & Li, 1998). There are also data mining techniques for clustering customers such as self-organizing-map (SOM—Kohonen, 1995) and techniques for discovering customer behavior such as association rules mining (Agrawal & Srikant, 1994). In the future, the availability of huge amounts of data about customers will compel marketers to adopt strong data mining tools to delve deep into customers' nature.

CONCLUSION

Equipped with advanced mobile technologies, more sophisticated marketing activities can be conducted now and in the future. In this article, we have discussed the benefits of mobile marketing, the role of mobile marketing in m-commerce, and the models used in mobile marketing. Although mobile marketing is powerful, it cannot replace other methods of marketing and should only be used as a powerful complement to traditional marketing. Mobile marketing should be integrated into the whole marketing strategy of a firm so that it can work seamlessly with other marketing approaches.

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very large Databases* (pp. 487-499), Santiago, Chile.

Ahonen, T. T. (2002). *M-profits: Making money from 3G services.* West Sussex, UK: John Wiley & Sons.

Ahonen, T. T., Kasper, T., & Melkko, S. (2004). *3G marketing: Communities and strategic partnerships.* West Sussex, UK: John Wiley & Sons.

Bayne, K. M. (2002). *Marketing without wires: Targeting promotions and advertising to mobile device users.* New York: John Wiley & Sons.

BusinessWeek. (2005). Special advertising section: 3G the mobile opportunity. *BusinessWeek* (Asian ed.), (November 21), 92-96.

Choon, S. L., Hyung, S. S., & Kim, D. S. (2004). A classification of mobile business models and its applications. *Industrial Management & Data Systems, 104*(1), 78-87.

Coursaris, C., & Hassanein, K. (2002). Understanding m-commerce. *Quarterly Journal of Electronic Commerce, 3*(3), 247-271.

Dey, A. K., & Abowd, G. D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction, 16*(2-4), 97-166.

Haig, M. (2002). *Mobile marketing: The message revolution.* London: Kogan Page.

Halonen, T., Romero, J., & Melero, J. (2003). *GSM, GPRS and EDGE performance: Evolution towards 3G/UMTS.* West Sussex, UK: John Wiley & Sons.

Holma, H., & Toskala, A. (2002). *WCDMA for UMTS* (2nd ed.). West Sussex, UK: John Wiley & Sons.

Jurvetson, S., & Draper, T. (1997). *Viral marketing.* Retrieved from http://www.dfj.com/cgi-bin/artman/publish/steve_tim_may97.shtml

Kohonen, T. (1995). *Self-organizing maps.* Berlin: Springer-Verlag.

Kwon, O.B., & Sadeh, N. (2004). Applying case-based reasoning and multi-agent intelligent system to context-aware comparative shopping. *Decision Support Systems, 37*(2), 199-213.

Ling, C. X., & Li, C.-H. (1998). Data mining for direct marketing: Problems and solutions. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 73-79), New York.

Linner, J. (2003). *Hitting the mark with text messaging.* Retrieved from http://wireless.sys-con.com/read/41316.htm

Scharl, A., Dickinger, A., & Murphy, J. (2005). Diffusion and success factors of mobile marketing. *Electronic Commerce Research and Applications, 4,* 159-173.

Tarasewich, P., Nickerson, R.C., & Warkentin, M. (2002). Issues in mobile e-commerce. *Communications of the Association for Information Systems, 8,* 41-84.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7,* 185-198.

Zhang, D. (2003). Delivery of personalized and adaptive content to mobile devices: A framework and enabling technology. *Communications of the Association for Information Systems, 12,* 183-202.

KEY TERMS

Bluetooth: Used mostly to connect personal devices wirelessly like PDAs, mobile phones, laptops, PCs, printers, and digital cameras.

Code Division Multiple Access (CDMA): A kind of 2G technology that allows users to share a channel by encoding data with channel-specified code and by making use of the constructive interference properties of the transmission medium.

Enhanced Data rates for GSM Evolution (EDGE): A kind of 2.5G technology. A new modulation scheme is implemented in EDGE to enable transmission speed of up to 384 kbps within the existing GSM network.

General Packet Radio Service (GPRS): Belongs to the family of 2.5G. GPRS is the first implementation of packet switching technology within GSM. The speed of GPRS can reach up to 115 Kbps.

Global System for Mobile (GSM) Communications: One of the 2G wireless mobile network technologies and the most widely used today. It can now operate in the 900 MHz, 1,800 MHz, and 1,900 MHz bands.

3G: The third generation of mobile telecommunication technologies. 3G refers to the next generation of mobile networks which operate at frequencies as high as 2.1 GHz, or even higher. The transmission speeds of 3G mobile wireless networks are believed to be able to reach up to 2 Mbps.

Time Division Multiple Access (TDMA): Divides each network channel into different time slots in order to allow several users to share the channel.

Time Division Synchronous Code Division Multiple Access (TD-SCDMA): A 3G mobile telecommunications standard developed in China.

2G: The second generation of mobile telecommunication technologies. It refers to mobile wireless networks and services that use digital technology. 2G wireless networks support data services.

2.5G: The second-and-a-half generation of mobile telecommunication technologies. 2.5G wireless system is built on top of a 2G network. 2.5G networks have the ability to conduct packet switching in addition to circuit switching. 2.5G supports higher transmission speeds compared to 2G systems.

W-CDMA: Developed by NTT DoCoMo as the air interface for its 3G network called FOMA. It is now accepted as a part of the IMT-2000 family of 3G standards.

Wireless Local Area Network (WLAN): Connects users wirelessly instead of using cables. WLAN is not a kind of mobile telecommunication technology. The coverage of WLAN may vary from a single meeting room to an entire building of a company.

Cache Invalidation in a Mobile Environment

Say Ying Lim
Monash University, Australia

INTRODUCTION

The rapid development, as well as recent advances in wireless network technologies, has led to the development of the concept of mobile computing. A mobile computing environment enables mobile users to query databases from their mobile devices over the wireless communication channels (Cai & Tan, 1999). The potential market for mobile computing applications is projected to increase over time by the currently increasingly mobile world, which enables a user to satisfy their needs by having the ability to access information anywhere, anytime. However, the typical nature of a mobile environment includes low bandwidth and low reliability of wireless channels, which causes frequent disconnection to the mobile users. Often, mobile devices are associated with low memory storage and low power computation and with a limited power supply (Myers & Beigl, 2003). Thus, for mobile computing to be widely deployed, it is important to cope with the current limitation of power conservation and low bandwidth of the wireless channel. These two issues create a great challenge for fellow researchers in the area of mobile computing.

By introducing data caching into the mobile environment, it is believed to be a very useful and effective method in conserving bandwidth and power consumptions. This is because, when the data item is cached, the mobile user can avoid requests for the same data if the data are valid. And this would lead to reduced transmissions, which implies better utilization of the nature of the wireless channel of limited bandwidth. The cached data are able to support disconnected or intermitted connected operations as well. In addition, this also leads to cost reduction if the billing is per KB data transfer (Lai, Tari, & Bertok, 2003). Caching has emerged as a fundamental technique especially in distributed systems, as it not only helps reduce communication costs but also offloads shared database servers. Generally, caching in a mobile environment is complicated by the fact that the caches need to be kept consistent at all time.

In this article, we describe the use of caching that allows coping with the characteristics of the mobile environment. We concentrate particularly on cache invalidation strategy, which is basically a type of caching strategy that is used to ensure that the data items that are cached in the mobile client are consistent in comparison to the ones that are stored on the server.

BACKGROUND

Caching at the mobile client helps in relieving the low bandwidth constraints imposed in the mobile environment (Kara & Edwards, 2003). Without the ability to cache data, there will be increased communication in the remote servers for data and this eventually leads to increased cost and, with the nature of an environment that is vulnerable to frequent disconnection, may also lead to higher costs (Leong & Si, 1997). However, the frequent disconnection and the mobility of clients complicate the issue of keeping the cache consistent with those that are stored in the servers (Chand, Joshi, & Misra, 2004).

Thus, when caching is used, ensuring data consistency is an important issue that needs considerable attention at all times (Lao, Tari, & Bertok, 2003). This is because the data that has been cached may have been outdated and no longer valid in comparison to the data from the corresponding servers or broadcast channel.

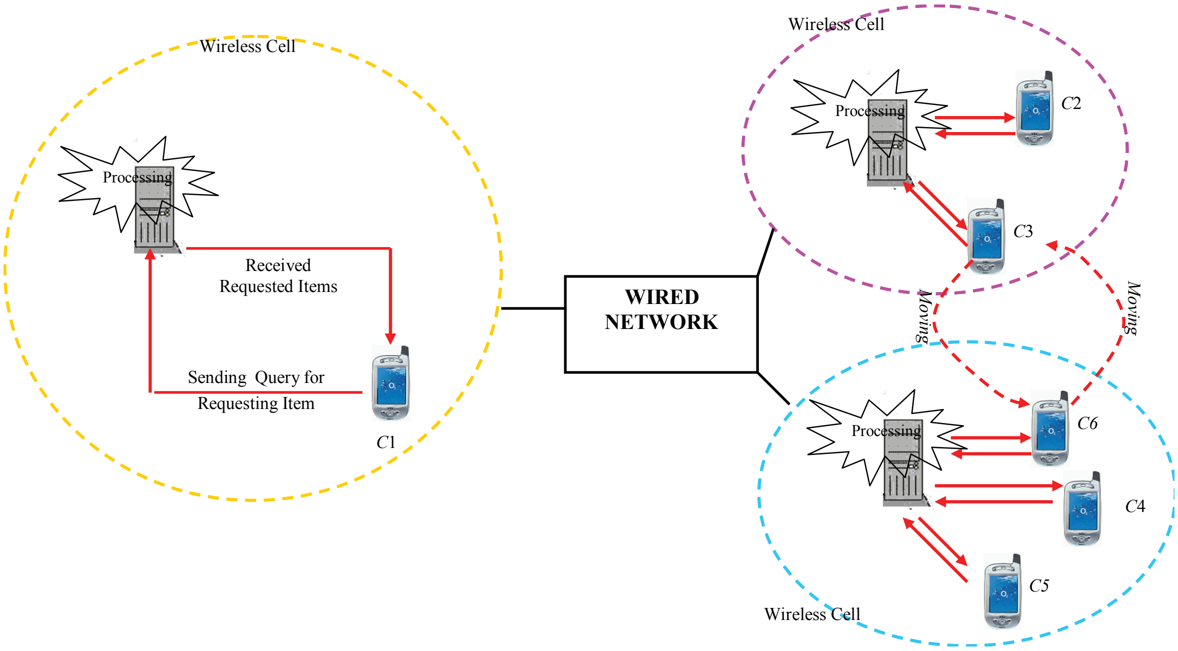
Figure 1 shows an illustration of a typical mobile environment that consists of mobile clients and servers, which are also know as mobile host (MH) and mobile support system (MSS) respectively. The mobile clients and servers communicate via a wireless channel within a certain coverage, known as cell (Chand, Joshi, & Misra, 2003; Cai & Tan, 1999). There are two approaches for sending a query in a mobile environment, which are: (a) The mobile clients are free to request data directly from the server via the wireless channel and the server will process and pass the desired data items back and (b) the mobile clients can tune into the broadcast channel to obtain the desired data items and download it to his/her mobile device. This can be illustrated in Figure 1a and Figure 1b respectively. The assumption is that updates are only able to occur at the server side and mobile clients can only have a read only feature.

CACHE INVALIDATION

Due to the important issue in the mobile environment, which is the ability to maintain data consistency, cache invalidation strategy is of utmost significance to ensure that the data items cached in the mobile client are consistent with those that are stored on the server. In order to ensure that data that are about to be used is consistent, a client must validate its cache prior to using any data from it.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Mobile environment architecture



There are several distinctive and significant benefits that cache invalidation brings to a mobile computing environment. If cache data are not validated to check for consistency, it will become useless and out-of-date. However, if one can utilize the cache data then the benefits it may bring include energy savings—that is, by reducing the amount of data transfer—and in return result in cost savings.

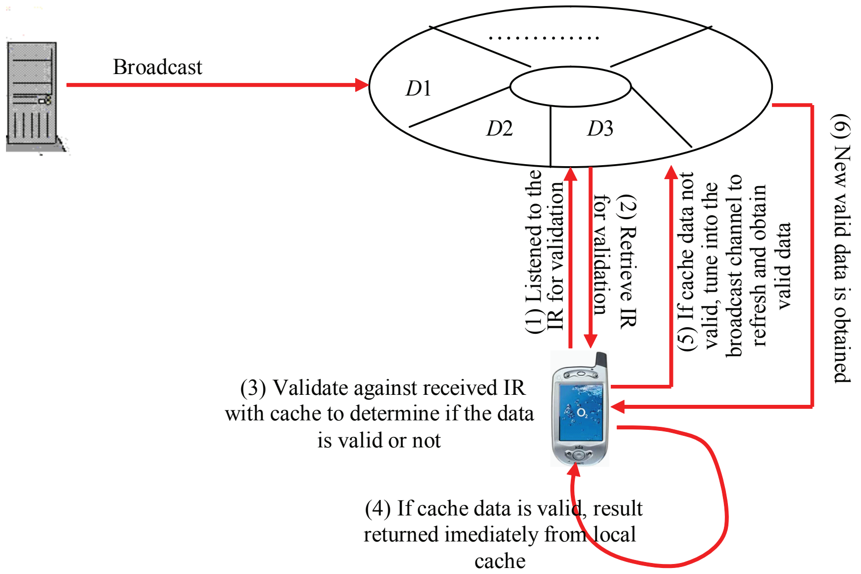
Using Cache Invalidation in a Mobile Environment

This can be done by using the broadcasting concept in communicating cache validation information to mobile clients. The server broadcasts the cache information, which is known as cache invalidation report (IR), periodically on the air to help clients validate their cache to ensure they are still consistent and can be used. It appears that the broadcast mechanism is more appropriate for the mobile environment due to its characteristic of salability, which allows it to broadcast data to an arbitrary number of clients who can listen to the broadcast channel anytime (Lai, Tari, & Bertok, 2003). By using the broadcasting approach, whereby the server periodically broadcasts the IR to indicate the change data items, it eliminates the need to query directly to the server for a validation cache copies. The mobile clients would be able to listen to the broadcast channel on the IR and use them to validate their local cache respectively (Cao, 2002).

Although cache invalidation strategy is important in a mobile environment, it will be vulnerable to disconnection and the mobility of the clients. One of the main reasons that cause mobile clients frequent disconnection is the limited battery power, and that is why mobile clients often disconnect to conserve battery power. It may appear to be very expensive at times to validate the cache for clients that experience frequent disconnection, especially with narrow wireless links. Other drawbacks would include long query latency, which is associated with the need of the mobile client to listen to the channel for the next IR first before he is able to conclude whether the cache is valid or not before answering a query. Another major drawback is the unnecessary data items in the IR that the server keeps. This refers to data items that are not cached by any mobile clients. This is thereby wasting a significant amount of wireless bandwidth.

Example 1: A mobile client in a shopping complex denoted as *C*1 in Figure 2 wanted to know which store to visit by obtaining a store directory. The client has previously visited this store and already has a copy of the result in his cache. In order to answer a query, the client will listen to the IR that are broadcasted and use it for validation against its local cache to see if it is valid or not. If there is a valid cached copy that can be used in answering the query, which is getting the store directories, then the result will be returned immediately. Otherwise, if the store directories have changed and now contain new shops, then the invalid caches have to be refreshed via sending a query to the server (Elmagarmid et al., 2003). The server would keep track of the recently updated data and broadcast the up-to-date IR every now and

Figure 2. Using cache invalidation in a mobile environment


Broadcast
....
D1
D2
D3
(1) Listened to the IR for validation
(2) Retrieve IR for validation
(5) If cache data not valid, tune into the broadcast channel to refresh and obtain valid data
(6) New valid data is obtained
(3) Validate against received IR with cache to determine if the data is valid or not
(4) If cache data is valid, result returned imediately from local cache

then for the clients to tune in. This can be done either by sending a request directly to the server (pull-based system) or tune into the broadcast channel (push-based system). Figure 2 illustrates an example of a push-based system.

In summary, effective cache invalidation strategies must be developed to ensure consistency between cached data in the mobile environment and the original data that are stored on the server (Hu & Lee, 1998).

Designing Cache Invalidation Strategies

It is important to produce an effective cache invalidation strategy to maintain a high level of consistency between cached data in the mobile devices with those that are stored on the server. In general, there are three possible basic ways in designing the invalidation strategies that described as follows (Hu & Lee, 1998).

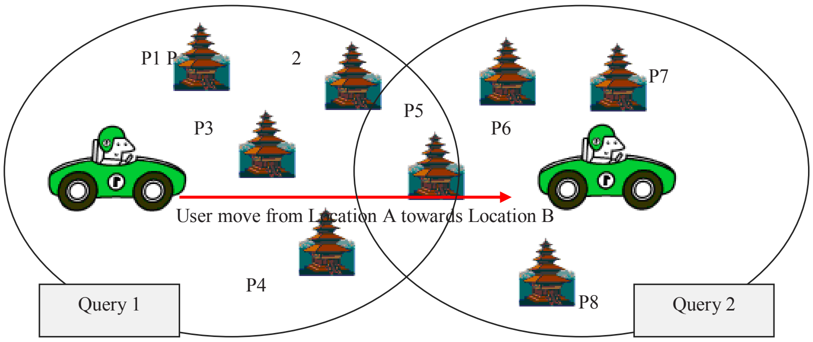
Assuming the server is stateful, whereby it knows which data are cached and by which particular mobile clients. Whenever there are changes in the data item in the server, the server would send a message to those clients, which has cached that particular item that has been updated or changed. In this way, the server would be required to locate the mobile clients. However, there is a major limitation in this method, that is, particularly in cases of disconnection. This is because mobile clients that are disconnected cannot be contacted by the server and thus its cache would have turned into invalid upon reconnection. Another aspect is if the mobile client moves to a new location, it will have to notify the server of the relocation. And all these issues, such as disconnection and mobility, have to be taken into account because it incurs costs from sending the messages to and from the server via the uplink and downlink messages.

The second possible way is to have the mobile client query the server directly in order to verify the validity of the cache data prior to using it. This appears to be straightforward and easy, but one has to bear in mind that this method would generate a lot of uplink traffic in the network.

In contrast to stateful method, another way that can be taken into account in designing the invalidation is using a stateless method. This method is in direct opposite from the first possible way, which is the stateful method. In this method, the server is not aware of the state of the client's cache and the client location and disconnected status. The server would not care about all these but just periodically broadcasts an IR containing the data items that have been updated or changed in comparison to its previous state. Thus the server just keeps track of which item is recently updated and broadcasts them in an IR. Then only the client determines whether its cache is valid or not by validating it against the IRs that are broadcasted on the wireless channel.

Another challenging issue that involves determining an efficient invalidation strategy is to optimize the organization of the IR. Commonly, a large-sized report provides more information and appears to be more effective. But publishing a large report also brings drawbacks, such as implying a long latency for mobile clients to listen to the report due to the low bandwidth wireless channel. There have been several methods proposed in addressing the report optimization issue in other works, such as using the dual report scheme and bit sequence scheme (Tan, Cai, & Ooi, 2001; Elmagarmid et al., 2003).

Figure 3. Architecture of location dependent query processing



Location-Dependent Cache Invalidation

Due to the fact that mobile users in a typical mobile environment move around frequently by changing location has opened up a new challenge of answering queries that is dependent on the current geographical coordinates of the users (Barbara, 1999; Waluyo, Srinivasan, & Taniar, 2005). This is known as location dependent queries (Kottkamp & Zukunft, 1998). In this location dependent query, the server would produce answers to a query based on the location of the mobile client issuing the query. Thus, a different location may sometimes yield a different result even though the query is taken from a similar source.

Figure 3 depicts an illustration of a location dependent query processing. This shows that when the mobile client is in Location *A*, the query would return a set of results and when the mobile client moves towards a new Location *B,* another set of results will be returned. However there are cases of results overlapping between nearby locations. An example of a location dependent query can be: “Find the nearest restaurants from where I am standing now.” This is an example of static object whereby restaurants are not moving. An example of a dynamic object would be: “What is the nearest taxi that will pass by me” (Lee et al., 2002).

With the frequent movement of mobile users, very often the mobile clients would query the same server to obtain results, or with the frequent movement of mobile users from location to location, very often the mobile clients would suffer from scarce bandwidth and frequent disconnection, especially when suddenly moved towards a secluded area (Jayaputera & Taniar, 2005). Hence, is essential to have data caching that can cope with cases of frequent disconnections. And often data may have become invalid after a certain point of time, especially in the area of location dependent.

Example 2: A mobile user who is in Location *A,* cached the results of the nearby vegetarian restaurants in Location *A*. As he moves to Location *B*, he would like another list of nearby vegetarian restaurants. The user is sending a query to the same source, but the results returned are different due to location dependent data. And because there are data previously cached, this data—which is the result obtained when he is in Location *A*—would become invalid since he has now moves to Location *B*.

Hence, location dependent cache invalidation serves the purpose of maintaining the validity of the cached data when the mobile client moves from one location to another location (Zheng, Xu, & Lee, 2002). The emergence of this location dependent cache invalidation is due to mobile client’s movement and thus the data value for a data item is actually dependent on the geographical location. Hence, traditional caching that does not consider geographical location is inefficient for location dependent data.

There are both advantages and disadvantages of location dependent cache invalidation. The major benefit that the attached invalidation information provided is that it provides a way for the client to be able to check for validity of cached data in respect to a certain location. These are necessary, especially in cases of when the mobile client wishes to issue the same query later when he/she moves to a new location. Another situation for the importance in checking the validity of the cached data is that because mobile clients keep on moving even right after they submit a query, they would have arrived to a new location when the results are returned. This may occur if there is a long delay in accessing data. Thus, if this two situations occur, then it is significant to validate the cached data because it may have become invalid (Zheng, Xu, & Lee, 2002).

FUTURE TRENDS

There have been several researches done in the area of exploring cache invalidation in a mobile environment. The usage of cache invalidation has obviously provoked extensive complicated issues. There are still many limitations of the nature of the mobile environment as well as mobility of the users that generate a lot of attention from research in finding a good cache strategy that can cope well with frequent disconnection and low power consumption.

In the future, it is critical to build an analytical model to get a better understanding of how cache invalidation works and how well it can cope in the mobile environment. Developing caching strategies that support cache invalidation for a multiple channel environment is also desirable, whereby a mixture of broadcast and point-to-point channels are being used. Including a dynamic clustering is also beneficial in order to allow the server to group data items together as their update changes. Besides these, further investigation on other cache replacement policies, as well as granularities issues, is also beneficial.

Due to the non-stop moving clients, further research on adapting cache invalidation into location dependent data is favorable. Another possible issue that could open up for future work may involve minimizing the waiting time for the mobile client in acquiring the IR, since the mobile client has to obtain an IR prior to their cache being validated. Thus, it is essential to be able to reduce waiting time. Another aspect is due to wireless channels that are often error prone due to their instability, bandwidth, and so on. Thereby, having techniques to handle errors in a mobile environment is definitely helpful. Last but not least, having further study on integrating several different strategies to obtain a more optimal solution in coping with mobile environment is advantageous.

CONCLUSION

Although there is a significant increase in the popularity of mobile computing, there are still several limitations that are inherent, be it the mobile device itself or the environment itself. These include limited battery power, storage, communication cost, and bandwidth problems. All these have become present challenges for researchers to address.

In this article, we have described the pros and cons of adopting cache invalidation in a mobile environment. We include adapting cache invalidation strategy in both location and non-location dependent queries. Discussion regarding the issue in designing cache invalidation is also provided in a preliminary stage. This article serves as a valuable starting point for those who wish to gain some introductory knowledge about the usefulness of cache invalidation.

REFERENCES

Barbara, D., & Imielinski, T. (1994). Sleepers and workaholics: Caching strategies in mobile environments. *MOBIDATA: An Interactive Journal of Mobile Computing*, *1*(1).

Cai, J., & Tan, K. L. (1999). Energy efficient selective cache invalidation. *Wireless Networks*, *5*(6), 489-502.

Chan, B. Y., Si, A., & Leong, H. V. (1998). Cache management for mobile databases: Design and evaluation. In *Proceedings of the International Conference on Data Engineering (ICDE)* (pp. 54-63).

Chand, N., Joshi, R., & Misra, M. (1996). Energy efficient cache invalidation in a disconnected mobile environment. In *Proceedings of the Twelfth International Conference on Data Engineering* (pp.336-343).

Cao, G. (2003). A scalable low-latency cache invalidation strategy for mobile environment. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, *15*(2), 1251-1265.

Cao, G. (2002). On improving the performance of cache invalidation in mobile environment. *Mobile Networks and Applications*, *7*(4), 291-303.

Deshpande, P. M., & Ramasamy, K. (1998). Caching multidimensional queries using chunks. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp. 259-270).

Dong Jung, Y. H., You, J., Lee, W., & Kim, K. (2002). Broadcasting and caching policies for location dependent queries in urban areas. In *Proceedings of the 2nd International Workshop on Mobile Commerce* (pp. 54-60).

Elmagarmid, A., Jing, J., Helal, A., & Lee, C. (2003). Scalable cache invalidation algorithms for mobile data access. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, *15*(6), 1498-1511.

Hu, Q., & Lee, D. (1998). Cache algorithms based on adaptive invalidation reports for mobile environment. *Cluster Computing*, pp. 39-48.

Hurson A.R., & Jiao, Y. (2005). Data broadcasting in mobile environment. In D. Katsaros, A. Nanopoulos, & Y. Manolopaulos (Eds.), *Wireless information highways* (Chapter 4). Hershey, PA: IRM Press.

Imielinski, T., & Badrinath, B. (1994). Mobile wireless computing: Challenges in data management. *Communications of the ACM, 37*(10), 18-28.

Jayaputera, J., & Taniar, D. (2005). Data retrieval for location-dependent queries in a multi-cell wireless environment. *Mobile Information Systems*, *1*(2), 91-108.

Kara, H., & Edwards, C. (2003). A caching architecture for content delivery to mobile devices. In *Proceedings of the 29th EUROMICRO Conference: New Waves in System Architecture (EUROMICRO'03)*.

Kottkamp, H.-E., & Zukunft, O. (1998). Location-aware query processing in mobile database systems. In *Proceedings of ACM Symposium on Applied Computing* (pp. 416-423).

Lai, K.Y., Tari, Z., & Bertok, P. (2003). Cost efficient broadcast based cache invalidation for mobile environment. In *Proceedings of the 2003 ACM symposium on Applied Computing* (pp. 871-877).

Lee, G., Lo, S-C., & Chen, A. L. P. (2002). Data allocation on wireless broadcast channels for efficient query processing. *IEEE Transactions on Computers*, *51*(10), 1237-1252.

Leong, H. V., & Si, A. (1997). Database caching over the air-storage. *The Computer Journal, 40*(7), 401-415.

Lee, D-L., Zhu, M., & Hu, H. (2005). When location-based services meet databases. *Mobile Information Systems*, *1*(2), 81-90.

Lee, D. K., Xu, J., Zheng, B., & Lee, W-C. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing*, *2*(3), 65-72.

Prabhajara, K., Hua, K. A., & Oh, J.H. (2000). Multi-level, multi-channel air cache designs for broadcasting in a mobile environment. In *Proceedings of the 16th International Conference on Data Engineering* (pp. 167-186).

Park, K., Song, M., & Hwang, C. S. (2004). An efficient data dissemination schemes for location dependent information services. In *Proceedings of the First International Conference on Distributed Computing and Internet Technology (ICDCIT 2004)* (Vol. 3347, pp.96-105). Springer-Verlag.

Tan, K. L., Cai, J., & Ooi, B. C. (2001). An evaluation of cache invalidation strategies in wireless environment. *IEEE Transactions on Parallel and Distributed Systems, 12*(8), 789-807.

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Research on location-dependent queries in mobile databases. *International Journal on Computer Systems: Science and Engineering*, *20*(3), 77-93.

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Research in mobile database query optimization and processing. *Mobile Information Systems*, *1*(4).

Xu, J., Hu, Q., Tang, X., & Lee, D. L. (2004). Performance analysis of location dependent cache invalidation scheme for mobile environments. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, *15*(2), 125-139.

Xu, J., Hu, Q., Lee, D. L., & Lee, W.-C. (2000). SAIU: An efficient cache replacement policy for wireless on-demand broadcasts. In *Proceedings of the 9th International Conference on Information and Knowledge Management* (pp. 46-53).

Xu, J., Hu, Q., Lee, W.-C., & Lee, D. L. (2004). Performance evaluation of an optimal cache replacement policy for wireless data dissemination. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, *16*(1), 125-139.

Yajima, E., Hara, T., Tsukamoto, M., & Nishio, S. (2001). Scheduling and caching strategies for correlated data in push-based information systems. *ACM SIGAPP Applied Computing Review*, *9*(1), 22-28.

Zheng, B., Xu, J., & Lee, D.L. (2002, October). Cache invalidation and replacement strategies for location-dependent data in mobile environments. *IEEE Transactions on Computers*, *51*(10), 1141-1153.

KEY TERMS

Caching: Techniques of temporarily storing frequently accessed data designed to reduce network transfers and therefore increase speed of download

Cache Invalidation Strategy: A type of caching strategy that is used to ensure that the data items that are cached in the mobile client are consistent in comparison to the ones that are stored on the server.

Caching Management Strategy: A strategy that relates to how client manipulates the data that has been cached in an efficient and effective way by maintaining the data items in a client's local storage.

Invalidation Report (IR): An informative report in which the changed data items are indicated; it is used for mobile clients to validate against their cache data to check if it is still valid or not.

Location-Dependent Cache Invalidation: maintaining the validity of the cached data when the mobile client changes locations.

Mobile Environment: Refers to a set of database servers, which may or may not be collaborative with one another, that disseminate data via wireless channels to multiple mobile users.

Pull-Based Environment: Also known as an on demand system, which relates to techniques that enable the server to process request that are sent from mobile users.

Push-Based Environment: Also known as a broadcast system where the server would broadcast a set of data to the air for a population of mobile users to tune in for their required data.

Communicating Recommendations in a Service-Oriented Environment

Omar Khadeer Hussain
Curtin University of Technology, Australia

Elizabeth Chang
Curtin University of Technology, Australia

Farookh Khadeer Hussain
Curtin University of Technology, Australia

Tharam S. Dillon
University of Technology, Sydney, Australia

INTRODUCTION

The Australian and New Zealand Standard on Risk Management, AS/NZS 4360:2004 (Cooper, 2004), states that risk identification is the heart of risk management. Hence risk should be identified according to the context of the transaction in order to analyze and manage it better. Risk analysis is the science of evaluating risks resulting from past, current, anticipated, or future activities. The use of these evaluations includes providing information for determining regulatory actions to limit risk, and for educating the public concerning particular risk issues. Risk analysis is an interdisciplinary science that relies on laboratory studies, collection, and exposure of data and computer modeling.

Chan, Lee, Dillon, and Chang (2002) state that the advent of the Internet and its development has simplified the way transactions are carried out. It currently provides the user with numerous facilities which facilitate transaction process. This process evolved into what became known as e-commerce transactions. There are two types of architectures through which e-commerce transactions can be conducted. They are: (a) client-server business architecture, and (b) peer-to-peer business architecture.

In almost all cases, the amount of risk involved in a transaction is important to be understood or analyzed before a transaction is begun. This also applies to the transactions in the field of e-commerce and peer-to-peer business. In this article we will emphasize transactions carried out in the peer-to-peer business architecture style, as our aim is to analyze risk in such transactions carried out in a service-oriented environment.

Peer-to-peer (P2P) architecture is so called because each node has equivalent responsibilities (Leuf, 2002). This is a type of network in which each workstation or peer has equivalent capabilities and responsibilities. This differs from client/server architecture, in which some computers or central servers are dedicated to serving others. As mentioned by Oram (2001), the main difference between these two architectures is that in peer-to-peer architecture, the control is transferred back to the clients from the servers, and it is the responsibility of the clients to complete the transaction. Some of the characteristics of peer-to-peer or decentralized transactions are:

1. There is no server in this type of transaction between peers.
2. Peers interact with each other directly, rather than through a server, as compared to a centralized transaction where the authenticity can be checked.
3. Peers can forge or create multiple identities in a decentralized transaction, and there is no way of checking the identity claimed by the peer to be genuine or not.

The above properties clearly show that a decentralized transaction carries more risks and hence merits more detailed investigation. Similarly, in a service-oriented peer-to-peer financial transaction, there is the possibility of the trusted agent engaging in an untrustworthy manner and in other negative behavior at the buyer's expense, which would result in the loss of the buyer's resources. This possibility of failure and the degree of possible loss in the buyer's resource is termed as risk. Hence, risk analysis is an important factor in deciding whether to proceed in an interaction or not, as it helps to determine the likelihood of loss in the resources involved in the transaction.

Risk analysis by the trusting agent before initiating an interaction with a trusted agent can be done by:

- determining the possibility of failure of the interaction, and
- determining the possible consequences of failure of the interaction.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. The riskiness scale and its associated levels

Riskiness Levels	Magnitude of Risk	Riskiness Value	Star Rating
Unknown Risk	-	-1	Not Displayed
Totally Risky	91-100% of Risk	0	Not Displayed
Extremely Risky	71-90% of Risk	1	From ⯪ to ★
Largely Risky	70% of Risk	2	From ★⯪ to ★★
Risky	26-50% of Risk	3	From ★★⯪ to ★★★
Largely Unrisky	11-25% of Risk	4	From ★★★⯪ to ★★★★
Unrisky	0-10% of Risk	5	From ★★★★⯪ to ★★★★★

The trusting agent can determine the possibility of failure in interacting with a probable trusted agent either by:

a. considering its previous interaction history with the trusted agent, if any, in the context of its future interaction, or
b. soliciting recommendations for the trusted agent in the particular context of its future interaction, if it does not have any previous interaction history with it.

When the trusting agent solicits for recommendations about a trusted agent for a particular context, then it should consider replies from agents who have previous interaction history with the trusted agent in that particular context. The agents replying back with the recommendations are called the *recommending agents*. But it is possible that each recommending agent might give its recommendation in its own way, and as a result of that, it will be difficult for the trusting agent to interpret and understand what each element of the recommendations mean. Hence, a standard format for communicating recommendations is needed so that it is easier for the trusting agent to understand and assimilate them. Further, the trusting agent has to determine whether the recommendation communicated by the recommending agent is trustworthy or not before considering it.

In this article we propose a methodology by which the trusting agent classifies the recommendation according to its trustworthiness. We also define a standard format for communicating recommendations, so that it is easier for the trusting agent to interpret and understand them.

BACKGROUND

Security is the process of providing sheltered communication between two communicating agents (Singh & Liu, 2003; Chan et al., 2002). We define *risk* in a peer-to-peer service-oriented environment transaction as the likelihood that the transaction might not proceed as expected by the trusting agent in a given context and at a particular time once it begins resulting in the loss of money and the resources involved in it. The study of risk cannot be compared with the study of security, because securing a transaction does not mean that there will be no risk in personal damages and financial losses. Risk is a combination of:

a. the uncertainty of the outcome; and
b. the cost of the outcome when it occurs, usually the loss incurred.

Analyzing risk is important in e-commerce transactions, because there is a whole body of literature based on rational economics that argues that the decision to buy is based on the risk-adjusted cost-benefit analysis (Greenland, 2004). Thus it commands a central role in any discussion of e-commerce that is related to a transaction. Risk plays a central role in deciding whether to proceed with a transaction or not. It can broadly be classified as an attribute of decision making that reflects the variance of its possible outcomes.

Peer-to-peer architecture-type transactions are being described as the next generation of the Internet (Orlowska, 2004). Architectures have been proposed by researchers (Qu & Nejdl, 2004; Schmidt & Parashar, 2004; Schuler, Weber, Schuldt, & Schek, 2004) for integrating Web services with peer-to-peer communicating agents like Gnutella. However, as discussed earlier, peer-to-peer-type transactions suffer from some disadvantages, and risk associated in the transactions is one of them. Hence, this disadvantage has to be overcome so that they can be used effectively with whatever service they are being integrated with.

Through the above discussion, it is evident that risk analysis is necessary when a transaction is being conducted in a peer-to-peer architecture environment. As mentioned before, risk analysis by the trusting agent can be done by determining the possibility of failure and the possible consequences of failure in interacting with a probable trusted agent. In order for the trusting agent to determine and quantify the possibility of failure of an interaction, we define the term riskiness. Riskiness is defined as the numerical value that is assigned by the trusting agent to the trusted agent after the interaction, which shows the level of possibility of failure of an interaction on the riskiness scale. The numerical value corresponds to a level on the riskiness scale, which gives an indication to other agents about the level of possibility of failure in interacting with a particular trusted agent. The riskiness scale as shown in Figure 1 depicts different levels of possibility of failure that could be present in an interaction.

The riskiness value to the trusted agent is assigned by the trusting agent after assessing the level of un-commitment in its actual behavior with respect to the promised commitment. The promised commitment is the expected behavior by which the trusted agent was supposed to behave in the interaction. The expected behavior is defined by the trusting agent according to its criteria, before starting its interaction with the trusted agent. The actual behavior is the actual commitment that the trusted agent showed or behaved in the interaction. Criteria are defined as the set of factors or bases that the trusting agent wants in the interaction and later against which it determines the un-committed behavior of the trusted agent in the interaction.

If the trusting agent has interacted previously with the trusted agent in the same context as its future interaction, then it can determine the possibility of failure in interacting with it by analyzing the riskiness value that it assigned to the trusted agent in their previous interaction. If a trusting agent has not interacted previously with a trusted agent in a particular context, then it can determine the possibility of failure in their future interaction, by soliciting for its recommendation from other agents who have dealt with the same trusted agent previously in the same context as that of the trusting agent's future interaction. As mentioned earlier the agents giving recommendations are called *recommending agents.*

But it would be difficult for the trusting agent to assimilate the data that it gets from the recommending agents and draw a conclusion if each agent gives its recommendation in its own format. It would rather be easier for the trusting agent if the recommendations came in a standard set or format that enables the trusting agent to ascertain the meaning of each element in the recommendations.

But even in the same context, each recommending agent might have different criteria in its interaction with the trusted agent. Consequently the riskiness value that it recommends for the trusted agent depends on its assessment of un-commitment in the trusted agent's actual behavior with respect to its expected behavior in those criteria. It would be baseless for the trusting agent to consider recommendations for a trusted agent in criteria of assessment which are not similar to those in its future interaction with that particular trusted agent. Additionally it is highly unlikely that the recommendations provided by the recommending agents would be completely reliable or trustworthy. Some agents might be communicating un-trustworthy recommendations. The trusting agent has to consider all these scenarios before it assimilates the recommendations from the recommending agents to assess the risk in dealing with a trusted agent.

In order to propose a solution to these issues, in the next sections we will define a methodology by which the trusting agent can classify the recommendations according to its trustworthiness. We also define a standard format for communicating recommendations so that the trusting agent can ascertain the meaning of each element of the recommendations before assimilating them, and consider only those whose criteria are of interest to it in its future interaction.

CLASSIFYING THE RECOMMENDATIONS AS TRUSTWORTHY OR UN-TRUSTWORTHY

As stated earlier, it is possible that the recommendation communicated by a recommending agent might not be trustworthy. The recommending agent might be communicating recommendations that the trusting agent finds to be incorrect or misleading after its interaction with the trusted agent. So the trusting agent has to determine whether the recommendation is trustworthy or not before assimilating it. To achieve that, we propose that each recommending agent is assigned a *riskiness value* while giving recommendations called *riskiness of the recommending agent (RRP).*

The riskiness value of the recommending agent is determined by the difference between:

- the riskiness value that the trusting agent found out for the trusted agent after interacting with it, and
- the riskiness value that the recommending agent recommend for the trusted agent to the trusting agent when solicited for.

When the trusting agent broadcasts a query soliciting for recommendations about a trusted agent in a particular context, it will consider replies from those agents who have interacted with that particular trusted agent previously in that same context. Hence, whatever riskiness value the recommending agents recommend to the trusting agent will be greater than -1, as -1 on the riskiness scale represents the riskiness value as *Unknown Risk,* which cannot be assigned to any agent after an interaction. After an interaction a value only within the range of (0, 5) on the riskiness scale can be assigned. So, the maximum range for the riskiness value of the recommending agent (RRP) is between (-5, 5), since this is the maximum possible range of difference between the riskiness value that the trusting agent might determine for the trusted agent after its interaction with it and the riskiness value recommended by the recommending agent for the trusted agent to the trusting agent.

We adopt the approach mentioned by Chang, Dillon, and Hussain (2006) which states that a recommending agent is said to be communicating trustworthy recommendations if its riskiness value while giving recommendations (RRP) is in the range of (-1, 1). A value within this range will state that there is a difference of one level in the riskiness value that the trusting agent found out after the interaction and what the recommending agent suggested for the trusted agent. If the riskiness value of the recommending agent is beyond those levels, then it hints that the recommending agent is giving recommendations that the trusting agent finds to vary a lot after the interaction, and there is at least a difference of two levels on the riskiness scale between what the trusting agent found and what the recommending agent recommended. An agent whose Riskiness value while giving recommendation (RRP) is beyond the level of (-1, 1) is said to be an *Un-trustworthy* recommending agent. Chang et al. (2006) mention that the trusting agent should only consider recommendations from agents who are either *Trustworthy* or *Unknown* in giving recommendations and leave the recommendation from agents who are *Un-trustworthy* in giving them. Hence the recommendation from agents with riskiness values beyond the levels of (-1, 1) will not be considered.

If the recommending agent gives more than one recommendation in an interaction, then its riskiness value while giving recommendation can be determined by taking the average of the difference of each recommendation.

Hence riskiness of the recommending agent (RRP) =

$$\frac{1}{N}\sum_{i=1}^{N} (\mathrm{Ti} - \mathrm{Ri})$$

where Ti is the riskiness value found out by the trusting agent after the interaction, Ri is the riskiness value recommended by the recommending agent for the trusted agent, and

N is the number of recommendations given by a particular agent.

DEFINING A STANDARD FORMAT FOR COMMUNICATING RECOMMENDATIONS

Whenever a trusting agent interacts with a trusted agent, a risk relationship is formed between them. The risk relationship consists of a number of factors. These include the trusting agent:

1. considering its previous experience with the trusted agent in the context of its future interaction with it, or soliciting recommendations for the trusted agent in the context of its future interaction if they have not interacted before;
2. determining the riskiness value of the trusted agent according to its previous interactions or recommendations;
3. predicting the future riskiness value of the trusted agent, within the time period of its interaction with the trusted agent; or
4. taking into consideration the cost of the interaction and assigning a riskiness value to the trusted agent after completing the interaction.

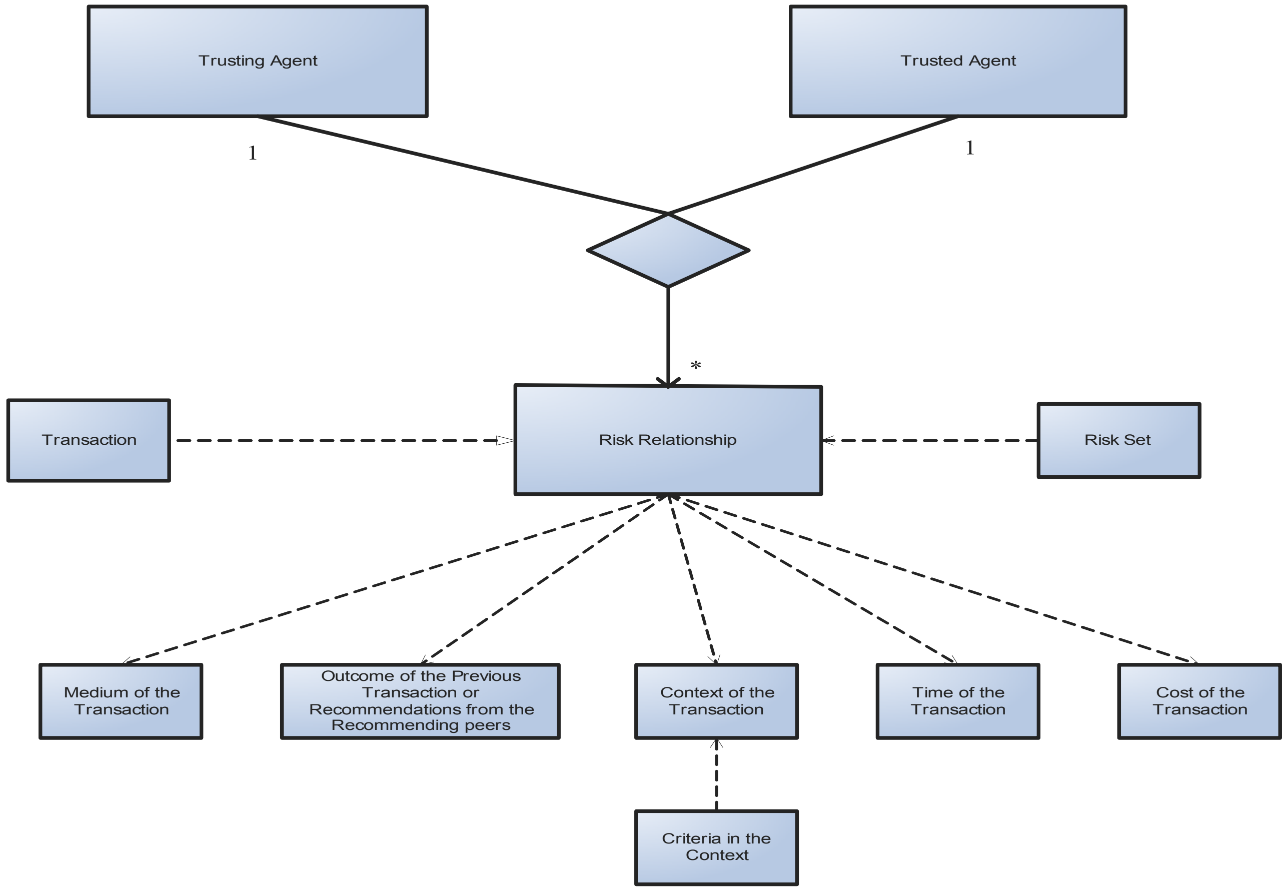
A risk relationship exists between a trusting agent and a trusted agent only if they interact with each other. Between a trusting agent and a trusted agent, there might exist one or more risk relationships depending on the number of times they interact with each other. For each interaction a new risk relationship is formed. Hence, the trusting agent and the trusted agent are in a ternary association (Eriksson & Penker, 2000) with the risk relationship as shown in Figure 2. But the risk relationship exists only if the trusting agent interacts with the trusted agent, and hence it is realized by a transaction between them. The risk relationship in turn is dependent on a number of factors. Figure 2 shows the risk relationship and the factors on which it is dependent.

As mentioned earlier, the trusting agent will consider recommendations from agents who have previous interaction history with the particular trusted agent in question in context similar to that of its future interaction with the trusted agent. A recommending agent when solicited for recommendation by a trusting agent for a particular trusted agent in a particular context will give its recommendation depending on its previous interaction with the particular trusted agent in the particular context. In other terms, it gives the risk relationship that it had formed with the trusted agent in that particular interaction as its recommendation. We propose that when any trusting agent solicits for recommendations for a trusted agent, then the recommending agents should give their replies in a standard format so that it is easier for the trusting agent to interpret their recommendation. The standard format is represented by a *risk set.*

The risk set is formed from the risk relationship that the recommending agent had from the last time it interacted with the particular trusted agent. Alternately, a risk set exists between any two agents only if there is a risk relationship between them, and hence it is dependent on the risk relationship as shown in Figure 2.

Once the risk relationship between any two agents has been established, then a risk set can be defined. The risk set contains the same elements as that of the risk relationship but in an ordered way. The order of appearance of the

Figure 2. Risk relationship that exists between any two agents


Trusting Agent
Trusted Agent
1
1
*
Transaction
Risk Relationship
Risk Set
Medium of the Transaction
Outcome of the Previous Transaction or Recommendations from the Recommending peers
Context of the Transaction
Time of the Transaction
Cost of the Transaction
Criteria in the Context

elements of the risk set is: {TP1, TP2, Context, CR, R', (Criteria, Commitment level), R, Cost, Start time, End time, RRP}, where:

- **TP1** denotes the trusting agent in the interaction. This is also the recommending agent while communicating recommendations.
- **TP2** denotes the trusted agent in the interaction.
- **Context** represents the context of the interaction.
- **CR** represents the 'current riskiness' value of the trusted agent before its interaction with the recommending agent. This is achieved either by the previous interaction history between the recommending agent and the trusted agent in the same context or by soliciting recommendations for the trusted agent by the recommending agent before its interaction,
- **R'** shows the predicted riskiness value of the trusted agent as determined by the recommending agent within the time slot of its interaction.
- **(Criteria, Commitment Level)** shows the factors or bases that the recommending agent used in its interaction with the trusted agent to assign it a riskiness value. These criteria are necessary to mention while giving recommendations, so that a trusting agent who asks for recommendation knows the factors on which this particular trusted agent has assigned the recommended riskiness value and only considers those recommendations which are of interest to it according to its criteria. Commitment level specifies whether the particular criterion was fulfilled by the trusted agent or not. A value of either 0 or 1 assigned here is based on its commitment. A value of 0 signifies that the criterion was not fulfilled by the trusted agent according to the expected behavior, whereas a value of 1 signifies that the criterion was fulfilled according to the expected behavior.
- **R** is the riskiness value assigned by the recommending agent to the trusted agent after its interaction. As discussed earlier, the riskiness value is determined after the interaction by assessing the level of un-commitment in the trusted agent's actual behavior with respect to the expected behavior.

- **Cost** represents the cost of the interaction.
- **Start Time** is the time at which the recommending agent started the interaction with the trusted agent.
- **End Time** is the time at which the interaction of the recommending agent ended with the trusted agent.
- **RRP** is the riskiness value of the recommending agent while giving recommendations. This value determines whether the recommendation is trustworthy or not.

To highlight the advantages of communicating the recommendations in a standard format by risk set and the usefulness of its elements, let us consider a scenario in which Bob wants to interact with a logistic company 'LC'. The context of its interaction with the logistic company is to 'transport its goods' on April 15, 2006. Let us represent the context as 'Transport'. The goods are worth $1,500. The criteria put up by Bob in its interaction with the logistic company 'LC' are:

1. Packing the goods properly.
2. Pickup of the goods on time by the logistic company.
3. Delivering the goods to the correct address on time as promised.

For explanation sake the criteria in the interaction are represented by C1, C2, and C3 respectively. The trusting agent Bob does not have any previous dealings with the logistic company 'LC', and in order to analyze the risk before proceeding in a business transaction with it, Bob solicits for recommendations from other agents who have previously dealt with logistic company 'LC' in a context similar to that in this interaction. The agents who had interacted previously in the same context give their recommendations to Bob in the form of a risk set, which relates to their previous interactions with the trusted agent 'LC'. Let us suppose that Bob receives recommendations from agents 'A', 'B', 'C', and 'D' in the form of risk set.

The recommendation from agent 'A' is:

{Agent 'A', Logistic Company 'LC', Transport, 5, 5, ((C3, 1) (C1, 0), (C2, 1)), 5, $5000, 15/07/2005, 22/07/2005, 0.8}

Similarly, recommendation from agent 'B' is:

{Agent 'B', Logistic Company 'LC', Transport, 3, 3, ((C5, 1) (C6, 0)), 3, $1000, 1/02/2006, 22/02/2006, -1}

Recommendation from agent 'C' is:

{Agent 'C', Logistic Company 'LC', Transport, 2, 3, ((C1, 0) (C2, 0), (C3, 1)), 2, UNKNOWN, 01/04/2006, 03/04/2006, -2.5}

Recommendation from agent 'D' is:

{Agent 'D', Logistic Company 'LC', Transport, 4, 4, ((C1, 1) (C2, 1), (C3, 1)), 5, UNKNOWN, 07/04/2006, 10/04/2006, 0}

The properties to be followed while forming or representing the risk set are:

1. The elements should be represented in the same order as defined above.
2. Each element of the risk set is mandatorally to be defined except the element 'cost'.
3. Each criteria and its commitment level should be represented inside a single '(', ')' bracket, separated by a comma, so as to differentiate it from the other criteria and the elements of the risk set.
4. If the cost is not represented, then it should be written as UNKNOWN.
5. If the riskiness value of the recommending agent (RRP) is not known, then it should be represented as UNKNOWN.
6. The elements of the risk set should be separated by a comma ','.

From the above recommendations it can be seen that:

- The recommendation from the recommending agent 'A' is trustworthy and exactly according to the criteria of the trusting agent's future interaction with the trusted agent. But there is a huge gap in time between the recommending agent's interaction with the trusted agent and the future interaction of the trusting agent with the trusted agent.
- The recommendation from recommending agent 'B' is trustworthy, but the criteria of its recommendation does not match with those of the trusting agent's future interaction with the trusted agent and so it is baseless for it to consider this recommendation.
- The criteria of recommending agent 'C' is similar to those of the trusting agent, but the riskiness value of the recommending agent 'C' (RRP) is not within the range of (-1,1). So it can be concluded that this is an un-trustworthy recommendation and it will not be considered by the trusting agent.
- The recommendation from recommending agent 'D' is trustworthy and in the criterions that the trusting agent wants in its interaction.

Hence, as can be seen, the trusting agent, by making use of the risk set, can interpret the meaning of each element of the recommendation that would help it to understand the recommendation better and assimilate it easily.

The advantages of communicating the recommendations in the form of a risk set are:

1. The recommendations come in a standard format and it is easier for the trusting agent to understand them.
2. Even if the context of two interactions is the same, the criteria might differ considerably, and the riskiness value assigned to each interaction is in accordance with its corresponding criteria. Therefore, while giving recommendations, the recommending agent must specify the criteria apart from the context of the interaction. By doing so, the trusting agent who is soliciting for recommendations might know the exact criteria in which the trusted agent was assigned the riskiness value recommended by the recommending agent and consider only those recommendations that are of interest to it. The risk set communicates the criteria along with the recommendations and also specifies the commitment level of the trusted agent in those criteria.
3. The risk set specifies the riskiness value of the trusted agent as determined by recommending agent before starting an interaction with it (CR), the predicted future riskiness value of the trusted agent within the time space of its interaction (R'), and the actual riskiness value of the trusted agent determined by the recommending agent (R) after its interaction with it depending on the level of un-commitment in the actual behavior of the trusted agent with respect to the expected behavior.
4. The risk set specifies the time of the interaction between the recommending agent and the trusted agent. As defined in the literature, risk is dynamic and it keeps on changing. It is not possible for an agent to have the same impression of another agent that it had at a given point of time. Hence, the trusting agent should give more weight to those recommendations which are near to the time slot of its interaction as compared to the far recent ones while assimilating them. This is achieved by using the proposed risk set, which specifies the context along with the accessing criteria and the riskiness values of the trusted agent according to the time assigned, in an ordered way.

CONCLUSION

In this article we discussed the need to analyze risk that could be associated in a peer-to-peer financial transaction. Further we discussed how a trusting agent can assess the possible risk beforehand that could be present in interacting with a particular trusted agent. We proposed a methodology of classifying the recommendations according to its trustworthiness. Further we discussed the risk relationship that exists between a trusting agent and a trusted agent in the post-interaction phase, and ascertain the factors on which the risk relationship is dependent. From that relationship we defined the risk set, which is an ordered way of representing the details of the transaction between the agents. This risk set is utilized by the recommending agents while communicating recommendations to the trusting agents, so that it can be interpreted and understood easily.

REFERENCES

Chan, H., Lee, R., Dillon, T. S., & Chang, E. (2002). *E-commerce: Fundamentals and applications* (1st ed.). New York: John Wiley & Sons.

Chang, E., Dillon, T., & Hussain F. K. (2006). *Trust and reputation in service oriented environments* (1st ed.). New York: John Wiley & Sons.

Cooper, D. F. (2004). *The Australian and New Zealand standard on risk management, AS/NZS 4360:2004, tutorial notes: Broadleaf Capital International Pty Ltd.* Retrieved from http://www.broadleaf.com.au/tutorials/Tut_Standard.pdf

Eriksson, H., & Penker, M. (2000). *Business modeling with UML: Business patterns at work.* New York: John Wiley & Sons.

Greenland, S. (2004). Bounding analysis as an inadequately specified methodology. *Risk Analysis, 24*(5), 1085-1092.

Leuf, B. (2002). *Peer to peer collaboration & sharing on the Internet.* Pearson Education.

Oram, A. (2004). *Peer-to-peer: Harnessing the power of disruptive technologies.* Retrieved February 16, 2004, from http://www.oreilly.com/catalog/peertopeer/chapter/ch01.html

Orlowska, M. E. (2004). The next generation messaging technology—makes Web services effective. *Proceedings of the 6th Asia Pacific Web Conference* (pp. 13-19). Berlin/Heidelberg: Springer-Verlag.

Qu, C., & Nejdl, W. (2004). Interacting the Edutella/JXTA peer-to-peer network with Web services. *Proceedings of the International Symposium on Applications and the Internet (SAINT'04),* Tokyo, (pp. 67-73).

Schmidt, C., & Parashar, M. (2004). A peer-to-peer approach to Web service discovery. *World Wide Web Journal, 7*(2), 211-229.

Schuler, C., Weber, R., Schuldt, H., & Schek, H. (2004). Scalable peer-to-peer process management—the OSIRIS approach. *Proceedings of the 2nd International Conference on Web Services*, San Diego, (pp. 26-34).

Singh, A., & Liu, L. (2003). TrustMe: Anonymous management of trust relationships in decentralized P2P systems. *Proceedings of the 3rd IEEE International Conference on P2P Computing* (pp. 142-149), Linköping, Sweden.

KEY TERMS

Recommending Agent: An agent who gives its recommendation about a trusted agent to a trusting agent, when solicited for.

Risk Set: A standard format for giving recommendations by the recommending agents.

Riskiness Scale: A scale that represents different levels of risk that could be possible in an interaction.

Riskiness Value: A value that is assigned to the trusted agent by the trusting agent after its interaction with it. This value specifies a level of risk on the riskiness scale that the trusted agent deserves according to the level of un-committed behavior in its interaction with the trusting agent.

RRP: Stands for riskiness value of the recommending agent. This value is used to determine if the recommending agent is communicating trustworthy recommendations or not.

Trusted Agent: An agent with whom the trusting agent deals with and reposes its faith in.

Trusting Agent: An agent who controls the resources and interacts with another agent after reposing its faith in it.

Content Personalization for Mobile Interfaces

Spiridoula Koukia
University of Patras, Greece

Maria Rigou
University of Patras, Greece and
Research Academic Computer Technology Institute, Greece

Spiros Sirmakessis
Technological Institution of Messolongi and
Research Academic Computer Technology Institute, Greece

INTRODUCTION

The contribution of context information to content management is of great importance. The increase of storage capacity in mobile devices gives users the possibility to maintain large amounts of content to their phones. As a result, this amount of content is increasing at a high rate. Users are able to store a huge variety of content such as contacts, text messages, ring tones, logos, calendar events, and textual notes. Furthermore, the development of novel applications has created new types of content, which include images, videos, MMS (multi-media messaging), e-mail, music, play lists, audio clips, bookmarks, news and weather, chat, niche information services, travel and entertainment information, driving instructions, banking, and shopping (Schilit & Theimer, 1994; Schilit, Adams, & Want, 1994; Brown, 1996; Brown, Bovey, & Chen, 1997).

The fact that users should be able to store the content on their mobile phone and find the content they need without much effort results in the requirement of managing the content by organizing and annotating it. The purpose of information management is to aid users by offering a safe and easy way of retrieving the relevant content automatically, to minimize their effort and maximize their benefit (Sorvari et al., 2004).

The increasing amount of stored content in mobile devices and the limitations of physical mobile phone user interfaces introduce a usability challenge in content management. The physical mobile phone user interface will not change considerably. The physical display sizes will not increase since in the mobile devices the display already covers a large part of the surface area. Text input speed will not change much, as keyboard-based text input methods have been the most efficient way to reduce slowness. While information is necessary for many applications, the human brain is limited in terms of how much information it can process at one time. The problem of information management is more complex in mobile environments (Campbell & Tarasewich, 2004).

One way to reduce information overload and enhance content management is through the use of *context metadata.* Context metadata is information that describes the context in which a content item was created or received and can be used to aid users in searching, retrieving, and organizing the relevant content automatically. Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves (Dey, 2001). Some types of context are the *physical context,* such as time, location, and date; the *social context,* such as social group, friends, work, and home; and the *mental context,* which includes users' activities and feelings (Ryan, Pascoe, & Morse, 1997; Dey, Abowd, & Wood, 1998; Lucas, 2001).

By organizing and annotating the content, we develop a new way of managing it, while content management features are created to face efficiently the usability challenge. Context metadata helps the user find the content he needs by enabling single and multi-criteria searches (e.g., find photos taken in Paris last year), example-based searches (e.g., find all the video clips recorded in the same location as the selected video clip), and automatic content organization for efficient browsing (e.g., location-based content view, where the content is arranged hierarchically based on the content capture location and information about the hierarchical relationships of different locations).

DATE, TIME, LOCATION, AND PROXIMITY

While context can be characterized by a large number of different types of attributes, the contribution of context attributes to content management is of great importance. We focus on a small number of attributes, which are considered the most important in supporting content management and also have the most practical implementations in real products, such as date, time, location, and proximity (nearby Bluetooth devices). Bluetooth is a short-range wireless technology used

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

to create personal area networks among user mobile devices and with other nearby devices.

The first two attributes, date and time, are the most common in use in a wide range of applications. They are used to organize both digital and analog content, and offer an easy way of searching and retrieving the relevant content automatically. For example, many cameras automatically add the date and time to photographs. Furthermore, the location where content is created is another useful attribute for searching the content (e.g., home, workplace, summer cottage). Mobile devices give users the possibility to create content in many different locations. Users can associate the location with the equivalent content in order to add an attribute to it that will enable them to find it easier. Finally, proximity also plays an important role in content management, as nearby Bluetooth devices can provide information both in social and physical context. While each Bluetooth device can be uniquely identified, information can be provided on nearby people by identifying their mobile phones. An example for physical context is the case of a Bluetooth-based hands-free car kit that can be used to identify that the user is in a car.

USABILITY ISSUES AND PROBLEMS

The expansion of the dimension of context information in order to include location, as well as proximity context, can be of benefit to users while they are able to store, access, and share with others their own location-based information such as videos and photos, and feel the sense of community growing among them (Kasinen, 2003; Cheverist, Smith, Mitchell, Friday, & Davies, 2001). But when it comes to proximity to be included in context information, the problem of *privacy* emerges. It appears that users are willing to accept a loss of privacy when they take into account the benefits of receiving useful information, but they would like to control the release of private information (Ljungstrand, 2001; Ackerman, Darrel, & Weitzner, 2001).

While context metadata is attached to content, when users share content, they have to decide if they share all the metadata with the content or they filter out all or some part of them. The cost for memory and transmission of metadata, as it is textual information, is not an important factor to influence this decision. When the user receives location and proximity information attached to content, he or she may also find out where and with whom the creator of the content was when the content was created. As a result, both the location of the content creator and the location of nearby people are shared along with the content information. If this information is private, the sharing of it could be considered as a privacy violation. This violation may be 'multiplied' if the first recipient forwards the content and the metadata to other users.

However, users seem to be willing to share context metadata attached to content, as it would be convenient if context metadata were automatically available with the content (so that users do not have to add this information manually). Furthermore, it would be very helpful for the recipient if the received content was annotated with context metadata so that the recipient does not have to annotate it manually and be able to manage the content more easily. For example, in the case of image and video content, the filtering of context metadata such as location and people could be useless, since these same items appearing in the image or video can be identified visually from the image content itself.

But what is meaningful information to the end user? It seems that users want meaningful information, but they are not willing to put too much effort in creating it, unless this information is expected to be very useful. In the case of location, it would be difficult for users to type the name of the place and other attributes manually, since it would require their time and effort. Thus it would be important if meaningful context metadata, which include the required information, are automatically generated.

Proximity information also needs to be meaningful. In this way, meaningfulness is important when attaching information on nearby devices in the form of metadata. If the globally unique Bluetooth device address and the real name of the owner of the device could be connected, this functionality would give meaningful information to the user.

It is hard to determine which information is useful, while what is useful information in one situation might be totally useless in another. For example, when looking at photo albums, what is thought to be useful information varies a lot. When one is looking at family pictures taken recently, it is needless to write down the names of the people, since they were well known and discernable. But it is different looking at family pictures taken many years ago: the same people may not be that easily recognizable.

It appears that useful information depends on a user's location, what the information is used for, and in which time span. In order to create meaningful information, users need to put much effort into getting the data, organizing it, and annotating it with context metadata. Ways to minimize their effort and maximize their benefit should be developed.

CONCLUSION

The increasing amount of stored content in mobile devices and the limitations of physical mobile phone user interfaces introduce a usability challenge in content management. The efficient management of large amounts of data requires developing new ways of managing content. Stored data are used by applications which should express information in a sensible way, and offer users a simple and intuitive way of organizing, searching, and grouping this information. Inadequate design of user interface results in poor usability and makes an otherwise good application useless. Therefore, it is necessary to design and built context-aware applications.

Issues of usefulness and meaningfulness in utilizing context metadata need to be further investigated. Usefulness depends on the type of metadata. As far as location and proximity are concerned, it appears that the more time has passed since the recording of the data, the more accurate the information needs to be. Furthermore, in the case of location information, the closer to one's home or familiar places the data refers to, the more detailed the information needs to be. A main usability challenge is the creation of meaningful context metadata automatically, without users having to add this information manually. There exist many ways for automatic recording of information about a user's context, but the generated information is not always meaningful.

Another field that requires further research is privacy. It seems that users are willing to accept a loss of privacy, provided that the information they receive is useful and they have control over the release of private information. Content management provides users with a safe, easy-to-use, and automated way of organizing and managing their mobile content, as well as retrieving useful information efficiently.

REFERENCES

Ackerman, M., Darrel, T., & Weitzner, D. J. (2001). Privacy in context. *Human Computer Interaction, 16,* 167-176.

Brown, P. J. (1996). The stick-e document: A framework for creating context-aware applications. *IFIP Proceedings of Electronic Publishing '96,* Laxenburg, Austria, (pp. 259-272).

Brown, P. J., Bovey, J. D., & Chen, X. (1997). Context-aware applications: From the laboratory to the marketplace. *IEEE Personal Communications, 4*(5), 58-64.

Campbell, C., & Tarasewich, P. (2004). What can you say with only three pixels? *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction*, Glasgow, Scotland, (pp. 1-12).

Cheverist, K., Smith, G., Mitchell, K., Friday, A., & Davies, N. (2001). The role of shared context in supporting cooperation between city visitors. *Computers &Graphics, 25,* 555-562.

Dey, A. K., Abowd, G. D., & Wood, A. (1998). CyberDesk: A framework for providing self–integrating context-aware services. *Knowledge Based Systems, 11*(1), 3-13.

Dey, A. K. (2001). Understanding and using context. *Personal & Ubiquitous Computing, 5*(1), 4-7.

Kaasinen, E. (2003). User needs for location-aware mobile services. *Personal Ubiquitous Computing, 7,* 70-79.

Kim, H., Kim, J., Lee, Y., Chae, M., & Choi, Y. (2002). An empirical study of the use contexts and usability problems in mobile Internet. *Proceedings of the 35th Annual International Conference on System Sciences* (pp. 1767-1776).

Ljungstrand, P. (2001). Context-awareness and mobile phones. *Personal and Ubiquitous Computing, 5,* 58-61.

Lucas, P. (2001). Mobile devices and mobile data—issues of identity and reference. *Human-Computer Interaction, 16*(2), 323-336.

Ryan, N., Pascoe, J., & Morse, D. (1997). Enhanced reality fieldwork: The context-aware archaeological assistant. In V. Gaffney, M. v. Leusen, & S. Exxon (Eds.), *Computer applications in archaeology.*

Schilit, B., & Theimer, M. (1994). Disseminating active map information to mobile hosts. *IEEE Network, 8*(5), 22-32.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *IEEE Proceedings of the 1st International Workshop on Mobile Computing Systems and Applications*, Santa Cruz, CA, (pp. 85-90).

Sorvari, A., Jalkanen, J., Jokela, R., Black, A., Kolil, K., Moberg, M., & Keinonen, T. (2004). Usability issues in utilizing context metadata in content management of mobile devices. *Proceedings of the 3rd Nordic Conference on Human-Computer Interaction*, Tampere, Finland, (pp. 357-363).

KEY TERMS

Bluetooth: A short-range wireless technology used to create personal area networks among user devices and with other nearby devices.

Content Management: Ways of organizing and annotating content in order to retrieve and search it more efficiently.

Context: Any information that can be used to characterize the situation of an entity.

Context Metadata: Information that describes the context in which a content item was created or received.

Entity: A person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves.

Location: The place where content is created by the user.

Usability: The effectiveness, efficiency, and satisfaction with which users can achieve tasks in the environment of mobile devices.

Content Transformation Techniques

Ioannis Antonellis
Research Academic Computer Technology Institute, Greece
University of Patras, Greece

Christos Bouras
Research Academic Computer Technology Institute, Greece
University of Patras, Greece

Vassilis Poulopoulos
Research Academic Computer Technology Institute, Greece
University of Patras, Greece

INTRODUCTION

The expansion of the Web is enormous and, more and more, people everyday access its content trying to make their life easier and their informational level complete. One can realize that lately the advances in computers are such that many appliances exist in order to offer to its users the chance to access any type of information. The use of microcomputers, such as PDAs, laptops, palmtops, mobile phones and generally mobile devices, has lead to a situation where a way had to be found in order to offer to the users the same information as if they had a normal screen device. Almost all the mobile devices offer "Web-ready" functionality, but it seems that few of the Web sites are considering offering to the mobile users the opportunity to access their pages from the mobile devices.

On the one hand, the widespread use of mobile devices introduces a new big market and many chances for research and development. On the other hand, the use of small screen devices introduces a basic constraint both to the constructors of the devices and to the users: the small screen limitation. This is making difficult for the users to establish a mental model of the data, often leading to user disorientation and frustration (Albers & Kim, 2000). Many other restrictions have to be taken under consideration when using small devices, especially the low resolution, the amount of the memory, and the speed of the processor. Additionally, when using such devices the users are often in places with distractions of noise, interruptions and movement of the handheld device (Jameson et al., 1998).

Many companies exist in order to offer to the users of small screen devices the opportunity to access Web pages by doing syntactic translation (AvantoGo, DPWeb, Palmscape, and Eudora). Syntactic translation recodes the Web content in a rote manner, usually tag-for-tag or following some predefined templates or rules. This method seems to be successful especially for the devices that have graphical display. But, in order to achieve this, the Web pages are scaled down and small devices like mobile phone (very small screen and low resolution) are problematic. This happens because either the graphics are too small or the letters and links cannot be explored.

Another major problem of the use of small screen devices is that users often migrate from device to device during a day and they demand to be able to work in the same way whether they work on their personal computer or their mobile phone. This is the main issue that is going to be analyzed in this article: the way of migrating data from device to device without damaging the integrity of the data and without distracting the user.

Migration is the process of taking data originally designed for display on a large screen and transforming it to be viewed on the small screen (Jameson et al., 1998). The main techniques that exist and are used for data migration are direct migration, data modification, data suppression and data overview. The first one, direct migration, is a very simple. The data are sent directly to the small screen device and the user navigates to the data by scrolling horizontally and vertically on the page. The second method is more complicated and data is shortened and minimized in order to be viewable in a small screen device. Data suppression technique removes parts of the data and presents parts of them and the latest technique is based on the focus and context model (Spence, 2001).

All the aforementioned techniques are useful and any of them can be used efficiently for different types of data. This is a difficult part for the construction of the small screen devices. The constructors of the devices cannot include all the implementations of the techniques or, even if they do, the user has to be asked which one to choose or try the different implementations while viewing a source of data. The differences between the aforementioned techniques are focused on the quality of the information shown to the user and the range of information that is shown. This means that

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

in some techniques the quality of the information shown is high but the amount of information shown is quite poor. One can think that the quality of information is more important while another can think that the amount of information is more important. This is a question that cannot be answered simply. What we can safely note is that the answer depends on the type of data that we want to present to the users.

The rest of this article is structured as follows. In the next section we present the efforts of some companies that offer to the users of small screen devices the opportunity to access Web pages by doing syntactic translation. The first method of transforming information, its use, its advantages and disadvantages are presented to the third section. The fourth section presents the data modification technique and how it is implemented, and the fifth section the data suppression technique. The next section covers the issues concerning data overview technique and the last section presents a summarization and general overview of the techniques.

DIRECT MIGRATION TECHNIQUE

The most simple and most often used technique is the direct migration technique. It is used mostly for Web pages and its scope is to send to the users exactly the same data regardless of the device in use. The users are free to interact with the data and they are actually responsible for making themselves comfortable with the amount of data that they are presented. We cannot say that it is a user-centric technique but it is very easy to be implemented, very fast and does not require much effort either for machine or human. The main problem, which is actually a failure of the technique, is that it produces data that needs horizontal scrolling in order to be accessed and that way the user is much distracted.

Some additional techniques are used together with the data migration technique in order to reduce or remove the horizontal scrolling problem. The additional technique is mainly the wrapping technique, which removes the horizontal scrolling by putting the extra data under the main page that is shown to the small screen. The problem is not solved but it becomes minor, because it does not lessen the amount of data but transforms the horizontal scrolling to vertical.

Another additional technique requires duplicate creation of the data. It is used very often for Web sites and the method is creating two kinds of pages for the same data: one for large screen devices and one for small screen. Surely, this technique has major problems. One is that someone has to create two totally different pages for the same content. The other and more crucial problem is the size of the World Wide Web and the fact that almost nobody has made any effort to create two types of Web pages makes the technique difficult to be applied.

Research has shown that the users react better when they are confronting vertical scrolling rather than horizontal (Nielsen, 1999). However even vertical scrolling—generally any kind of scrolling—affects negatively the completion of any task (Albers & Kim, 2000; Dyson & Haselgrove, 2001; Jones et al., 1999). The above implies that this technique can be suitable only for situations where the user just wants to access and read some kind of information and the interaction level between the user and the data remains low.

Summarizing, we can say that this technique is very suitable for short text, sequential text, lists and menus that can be displayed within the width constraints of small screens (the impact of migration). It is not recommended to be used when the data include big tables and images (big, high resolution) because these types of data add horizontal scrolling that cannot be transformed.

DATA MODIFICATION

In this section we will analyze the second method for data migration, which is the data modification technique. Its main idea approaches the direct migration technique, but the data modification technique has countered the problem of big images and tables. When the data are to be presented to a small device, the size of the images, tables and lists is reduced and some parts of the text are summarized. In this way the users can save in download time and device memory (Mani, 2001).

The text summarization is the difficult part of the technique and it introduces a whole new theme for discussion. Many approaches have been proposed (Buyukkokten et al., 2000; Fukushima, 2001, Mani, 2001; Amitay & Paris, 2000). Some of them require a human expert to create the summaries while some others are based on machines.

The data that is presented to the users is a reduced form of the actual data. The user has the option to scroll vertically through the data that he comes up with. He can also select a part of the reduced data in order to "open" in another page of his small screen device the real text, which is hidden behind. This procedure can be algorithmic. When data are presented in this way to the user, then the procedure is to read the summarized, reduced data, select a specific topic that suits the user's needs, read the whole data that is hidden behind the summarized and then go back. The procedure then starts from the beginning.

We can say that this technique is very similar to the aforementioned direct migration technique but it goes one step further. It is used mainly for Web browsing where the data are already reduced and offer the user a style of navigation. The summarization that is included, whether it is for images (lower size, resolution) or text (summary), is very helpful for the end-user as it lessens the scrolling either vertical or horizontal. Actually this method does not have horizontal scrolling at all except for some specific, very rare conditions (very large images or tables).

Summarizing, we can say that this technique is very useful when users are determined of the information and can easily understand what they are looking for, from a summary of text or simple keywords. It cannot be useful for very specialized texts with difficult and mannered terminology. In general, the summaries have to be very specific and represent accurately the meaning of the text. The main problem of all the summarization techniques is that they do not succeed very often and cannot replace numerical data like financial information, weather information and dates. If one can think that some users want their small screen devices for accessing their bank accounts, watching the weather in a place that they visit or finding the financial exchange then this technique cannot be recommended.

DATA SUPPRESSION

As we are able to figure out from the name of this technique, what it actually does is to remove parts of the data that "seem" to be unimportant. What is presented to the user is the basic frame of the data. Displaying only skeleton information can simplify navigation and may reduce disorientation (Spence, 2001).

The data is not removed randomly, but there exist several techniques that help in this direction. Some methods for suppressing data is to select only some of the keywords (that are produce from text summarization), present only a specific number of words from each sentence or Z-thru mapping that imposes selective display (Spence, 2001).

This approach is very similar to the previous but it seems to be more compact. Very few data is presented to the user and most of the time there is no scrolling at all. The absence of scrolling has advantages and disadvantages. When there is no scrolling the user is not distracted from completing his task, but no scrolling means that the data is extremely reduced in order to fit the screen and it may be difficult for a simple user to locate the information he/she wants.

Navigating through the data in this technique is like a file system. The user has a list of words (like a folder) for each amount of data and by selecting an element of the list the information is expanded (files, subfolders) and shown to the user. Every time the user is able to return to the starting frame of data and start exploring from the beginning.

Like the previous, this technique has applications where the users know the exact information that they are looking for and they can figure it out from just a heading or a set of keywords. Searching through this type of data is almost impossible because the little amount of data that is presented is often not representative of the data that it comes from. However it is very useful for browsing through news portal when just a title or part of the title is enough for the user to understand the meaning of the whole article. It is used for structured data, which include information hierarchically structured. Sequential data with little or no structure could be less compatible to manipulate into categories for suppression (impact of migration).

DATA OVERVIEW

The last technique that is used for data migration is data overview technique. In reverse to the aforementioned techniques, which reduce parts of the data, this technique creates an overview of the whole data and presents it to the users. The whole data is minimized and the whole information is presented to the user minimized in order to fit the small screen of the device. It is based on the "focus and expand" method. When the user points a specific set of data that is contiguous then it is expanded and shown bigger to the screen in order to fit the screen and be readable.

The approach makes it easier for the users to access at once a very large amount of data without losing or not seeing any part of it and, in this way, the disorientation is lessened (Spence, 2001; Storey et al., 1999). Some methods that are used concerning the data overview technique are:

- Focus and context (Spence, 2001; Buyukkokten et al., 2000; Bjork, 2000)
- Fisheye Techniques (Spence, 2001; Storey et al., 1999)
- Zoom and pan (Good et al., 2002; Spence, 2001)
- Content lens (Dieberger et al., 2002)

In general, the technique seems to be problematic as the user is presented with a large amount of data in a small screen. The data is shrunk in order to fit the screen and may be difficult for the user even to see it and figure out what he is looking for. Movements while using a small screen device could create further distortion, or could make it difficult to discern what has been distorted (MacKay & Watters, 2003).

The nature of this approach produces both positive and negative points for the end-users. The point that the user is presented the whole information can be both positive and negative depending on the amount of data. However, it is very useful for the users to have full observation of the information they are looking for. The navigation is easy and is based at presenting in large the parts that are focused from the user, but the user can focus only on a part of information and he is not able to combine parts of data.

In general this method seems to be the best when the information that is accessed by the user includes large images, big tables, maps, graphs and in general everything that a "focus" method cannot distort but help.

OVERVIEW OF THE TECHNIQUES

In the previous sections we have discussed and analyzed the most common methods for data migration from large screen displays to small screen. As we can obviously see, each method has its advantages and disadvantages making difficult the selection of only one of them in order to cope with every type of data.

Direct migration cannot preserve scrolling and it is the fastest and easiest way to present data that are for reading. Its simplicity is its power but we can admit that is not user friendly.

Data modification technique solves many problems of the previous technique but still scrolling is an issue. At least paging of the data is preserved and the user can see a large part of the information in only one screen. The matter that rises from this technique is the summarization of the information, which may be distracting or not useful depending on the type of information. It could be seen as a good method for Web browsing.

Data suppression goes one step further than the previous technique by removing parts of data and summarizing the rest. It is named as the best method for browsing news portals where just the keywords of a news title can represent successfully the whole article. It is very weak for textual data and for searching, as it provides in a hierarchic manner only some keywords and often distracts a user that does not know exactly what he is looking for.

Data overview has a different angle of view than the three previous methods. It is based on the idea "focus and expand" and the philosophy is to present to the user all the information. When the data include large images, big tables and graphs, data overview is the best method for migrating data because it does not lessen or break into many pages all this information, which is by nature connected. On the contrary, when the user wants to read a text or browse in a big portal then this technique seems to be weak, as it provides to the user all the information in one screen and the data are often unreadable.

Summarizing, all the techniques offer to the users the opportunity to access any kind of information through their small screen devices like they would do to big screen ones. It is not fair to select one of them as the best one because each one is created for coping with different types of information and data. A device that could combine the implementation of all the aforementioned techniques could be a solution, but the complexity of modern life would prevent us to permute to the users the effort of data migration and thus make modern life more complex.

REFERENCES

Albers, M. J., & Kim, L. (2000). User Web browsing characteristics using Palm handhelds for information retrieval. In *Proceedings Of IPCC/SIGDOC Technology & Teamwork* (pp. 125-135). September, 2000, Cambridge, MA: IEEE.

Amitay, E., & Paris, C. (2000, November). Automatically summarizing Web sites: Is there a way around it? In *Proceedings of the 9th Internet Conference on Information and Knowledge Management* (pp. 173-179). McLean, VA.

AvantGo, Inc. (n.d.). *AvantGo*. Retrieved from http://www.avantgo.com

Bjork, S. (2000, May). Hierarchical flip zooming: Enabling parallel exploration of hierarchical visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 232-237). Palermo, Italy.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). *Seeing the whole in parts: Text summarization for Web browsing on handheld devices.* Retrieved from http://wwwconf.ecs.soton.ac.uk/archive/00000067/01/index.html

Dieberger, A., & Russell, D. M. (2002, January). Exploratory navigation in large multimedia documents using context lenses. In *Proceedings of 35th Hawaii International Conference on System Sciences* (pp. 1462-1468). Big Island, Hawaii.

Digital Paths LLC. DPWeb. Http://www.digitalpaths.com/prodserv/dpwebdx.htm

Dyson, M., & Haselgrove, M. (2001). The influence of reading, speed and line length and effectiveness of reading from screen. *International Journal Human Computer Studies, 54*(4), 585-612.

Fukusima, T., & Okumura, M. (2001, June). Text summarization challenge: Text summarization evaluation in Japan. In *Proceedings North American Association for Computational Linguistics* (pp. 51-59). ittsburgh, Philadelphia, Association of Computational Linguistics.

Good, L., Bederson, B., Stefik, M., & Baudisch, P. (2002). Automatic text reduction for changing size constraints. In *Proceedings of Conference on Human Factors in Computer Systems, Extended Abstracts* (pp. 798-799). April 2001, Minneapolis, MN.

ILINX, Inc. (n.d.). *Palmscape*. Retrieved from http://www.ilinx.co.jp/en/products/ps.html

Jameson A., Schafer, R., Weis, T., Berthold A., & Weyrath, T. (1998). Making systems sensitive to the user's time and working memory constraints. In *Proceedings of 4th international Conference on Intelligent User Interfaces* (pp. 79-86). December 1998, Los Angeles, CA: ACM Press.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving Web interaction on small displays. In *Proceedings of the 8th International WWW*

Conference. May 1999, Toronto, Canada. Retrieved from http://www8.org/w8-papers/1b-multimedia/improving/improving.html

MacKay, B., & Watters, C. (2003, Winter). The impact of migration of data to small screens on navigation. *IT&Society, 1*(3), 90-101.

Mani, I. (2001, October). Text summarization and question answering: Recent developments in text summarization. In *Proceedings of the 10th International Conference on Information and Knowledge Management* (pp. 529-531).

Nielsen, J. (1999, December). *Changes in usability since 1994*.

QUALCOMM, Inc. (n.d.). *Eudora Internet Suite*. Retrieved from www.eudora.com/internetsuite/eudoraweb.html

Spence, Robert. (2001). *Information visualization*. New York: *ACM Press*.

Storey, M. D., Fraachia, F., Davic, M., & Hausi, A. (1999, June). Customizing a fisheye view algorithm to preserve the mental map. *Journal of Visual Languages and Computing*, 254-267.

KEY TERMS

Content Transformation: The procedure that leads to changes to content in order to make it interoperable.

Migration: Migration is the process of taking data originally designed for display on a large screen and transforming it to be viewed on the small screen.

Small Screen Devices: Devices with small screen size where it is difficult to access large-sized blocks of information.

Syntactic Translation (of WWW Data): The recoding of the Web content in a rote manner, usually tag-for-tag or following some predefined templates or rules.

Context-Adaptive Mobile Systems

Christian Kaspar
University of Goettingen, Germany

Thomas Diekmann
University of Goettingen, Germany

Svenja Hagenhoff
University of Goettingen, Germany

INTRODUCTION

Even though a major part of the industrialized world works with computers on a daily basis and operating computers became much easier since the introduction of graphic interfaces, many users do not experience their computers as work relief, but rather as an increased burden in their everyday lives. One of the most important reasons for this attitude is the unnatural mode of communication between user and computer: the natural interpersonal communication takes information from the communication situation (e.g., the location of the interacting communicators, their personal preferences, or their relationship with each other) implicitly into account. On the other side, despite the development of new interfaces—such as voice and character recognition, which are much closer to interpersonal communication than keyboard terminals—communication between user and computer is still complex and characterized by little intuition. This is where the objectives of the context-adaptive systems come into play: it is the aim of context-adaptive systems to implicitly collect information about the situation of a system request (context) in order to enable more efficient communication between user and computer.

Currently, the concept of context-awareness and context-adaptation has attracted particular attention in the area of mobile communication. This is largely due to the fact that the obligatory requirement of the devices' portability leads to certain constraints of mobile devices. Small-sized screens, low data processing capacity, and inconvenient ways of navigation and data entry are some examples for these constraints. To overcome these limitations is particularly relevant in the area of multimedia Internet content and therefore requires the communication to be as efficient as possible. One possible option to reduce the resulting problem of presentation and selection of content on mobile devices is to automatically offer the user only those contents relevant for the concrete situation of the service request. Such services require that the computer can sense the particular situation of the service request and autonomously respond with appropriate actions.

AUTOMATED CONTEXT-AWARENESS

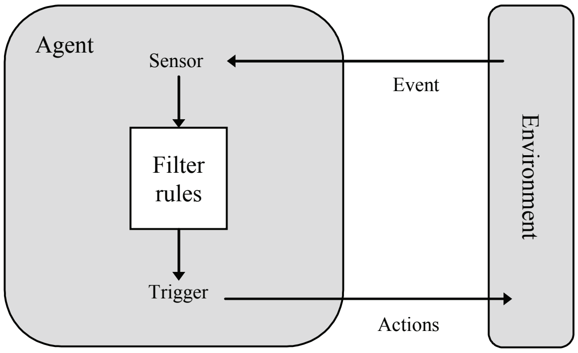
In order for real situations to be sensed automatically by computing devices, the situations of these system requests have to be computed as abstract, automatically understood events, so-called contexts. A context is any information that is used to characterize relevant situations of people, locations, or objects that are important for the interaction between application and user (Dey, 2001). A common classification of context information traces back to Schilit, Adams, and Want (1994). They distinguish between the technical context of participating and available computing resources, the social context of users that are involved in the system interaction, and the physical context of the location of the system interaction.

Computing context describes available network connections and network bandwidth. Additionally, computing context includes accessible peripherals such as printers, screens, or additional terminals. For example, if a multimedia application knows the user's available network bandwidth, it is able to adapt a video stream to its capacity and ensures streaming without jerks and with the highest possible resolution. Furthermore, if the multimedia application is aware of a high-resolution display close to the user, it can suggest this device as an alternative screen for displaying the video stream. To be able to identify each other, mobile computing devices must have radio or infrared sensors. A computing device equipped with radio or infrared sensors spans a distinct logical space (a so-called "smart space") within its sensor coverage. If a device enters another device's sensor space, the device will identify itself and send its network address or appropriate commands for application requests.

The social context contains information about the users involved in the interaction. The user's information, such as identity, age, gender, and preferences, can be gathered either explicitly from surveys or implicitly by observing the user's behavior. Surveying each user's personal characteristics and preferences is the most common form of gathering user information. Most often, surveying user information is directly linked with service registration. Because the provider

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Components of an agent system



has little means to control (usually voluntary) submitted information, information from user surveys is often of poor quality. Additionally, profiles that were gathered from a one-time survey remain static over time. Therefore, apart from voluntary authentication information on a specific Web site, the user can be additionally identified on the basis of his or her behavior. Every Web server has a protocol component that logs every server activity and stores these logs chronologically into different application-oriented protocol files. Analyzing these server protocols, it can be determined what requests for which resources have been completed during a specific unique Web site visit. To link recorded requests with an individual user, the IP address of the user's device or identification data stored as cookie on the user's device can be used.

Information about the physical context can be collected from a multitude of data sources such as contact-, thermo-, humidity-, acceleration-, torsion-, or photo-sensors, cameras, and microphones. Sensors that are equipped with processors not only collect data but also pre-process this data. Additionally, they can identify specific patterns such as fingerprints. The user's interaction location is of particular importance for the perception of the physical interaction context. "Location-based-services" are services that take the location of the user into consideration. These services promise to have great chances on the market (Lehner, 2003). The geographical location of a user that is required for services of this kind can either be determined by terminal-locating or by external network-locating.

Terminal-locating is carried out by an especially designed device that autonomously executes location measurements. The global positioning system (GPS) operated by the U.S. military is the best-known technology for self-locating. Receiving positioning signals that are beamed down by GPS satellites, a GPS device can accurately triangulate its position for up to 10 meters. Techniques that can position a location or object from a photo are more sophisticated than the GPS system. Yet they strongly resemble human orientation. Using photo cameras, these methods can calculate the angle and distance to a specific object (such as a building) with the help of a stored three-dimensional model.

Network-locating fixes a device's position using network information. The best-known method for network-locating is the cell identity technique (or cell of origin technique). This technique locates a mobile device within a cellular radio network using the network's cell-ID. Other network-locating techniques fix a particular position based on time differences of signals arriving at different base stations, the angle of arrival, or attenuation of signals from different base stations.

While Schilit et al. (1994) differentiated between three forms of context, Dey (2001) adds the primary and secondary context to these categories (Conlan, Power, & Barrett, 2003). The location, the type of device, the behavior of the user, and the time of inquiry represent primary request contexts. On the other hand, secondary request contexts are composed of a combination of primary context data. By taking the location and other people in the vicinity into consideration to form the secondary social context, the social situation of the user can be determined; for example, the user might be in his office with his colleagues, or out with friends. Furthermore, Chen and Kotz (2000) differentiate between the active and the passive context. The active context determines a change in behavior for the present application (e.g., by defining sections of interest for an adaptive online newspaper). A passive context shows the change in context conditions for the system inquiry only as extra information for the user. For example, a recommendation system can suggest a specific purchase in an online shop, or the user can see his location when using a navigation system.

CONTEXT-RELATED SYSTEM ADAPTION

A system is considered context-adaptive if it uses context information to offer its users relevant information or, rather, services (Dey & Abowd, 2000). Generally speaking, a context-adaptive system is characterized by a certain degree of autonomy when fulfilling its tasks. Therefore, adaptive systems are also referred to as agent systems (Russel & Norvig, 2003). Agents are software systems that, with the help of sensors, identify their environment as application-related events and by using pre-defined rules that activate respective events (see Figure 1). The actions that have been triggered by the agent can refer to information or they can contain control commands for other systems. The agent can either perform the action directly and autonomously (active context-awareness), or these actions are only a suggestion to the user (passive context-awareness).

A context-adaptive application consists of software objects that are automatically requested when the system senses a certain event in the systems environment. If the application identifies such a key event, it executes the respective object, or rather, the relevant methods that are part of this object. Usually, the adaptive application needs to combine raw data from its sensory perception of its environment to compute such a key event (e.g., GPS-coordinates, entries in a Web-server's log file, or identified, closely related network resources) since an individual piece of raw data is only able to hint at the relevant situation. An adaptive traffic information service, for instance, needs to compute information about the current date and time (system clock), the user's current position (e.g., GPS), the user's destination, and his or her preferred routes and means of transportation (user profile). In this context, we can distinguish different adaptation methods, depending on the type and extent of combined raw data collected while computing such a key event.

Raw sensor data that closely related to the system (such as time, temperature, or the device's data processing capacity)—whose values are, in accordance with the closed-world-assumption, known ex ante—are usually directly linked to the respective application-specific event. Therefore, dynamic key-value-pairs are composed, whereby the application-event contains the key and the sensor system contributes the key value (Chen & Kotz, 2000).

With regard to raw sensor data that cannot be directly linked to application events, such as user preferences or information about the location, however, the process is different. In this case, values from various sensor systems need to be aggregated. For instance, a standardized locating technique that is able to locate a user in closed rooms and outside does not yet exist. In order to locate users accurately inside and outside of buildings, data from different locating systems need to be combined. Since different locating systems utilize different measures (e.g., distance, angular separation, or geometrical position), these values need to be translated into standardized measures first and then have to be transferred into a joint data model (Hightower, Brumitt, & Borriello, 2002). This process is commonly known as "sensor fusion" (Chen, Li, & Kotz, 2004).

Adaptive systems that process contexts whose quality rating is subject to a high level of changeability (e.g., a user's movement in a room or outside of a building) need a high number of key values to be computed. Generally speaking, it does not make much sense to compose one key-value-pair for each of these values. Instead, artificial intelligence techniques, such as artificial neural networks that have been developed to process knowledge, can be utilized to compute these values. An artificial neural network (ANN) is a system that consists of a multitude of identical, networked computing elements, so-called neurons. ANNs are thus able to simultaneously process extensive and changeable raw data in an efficient manner (Van Laerhoven, Aidoo, & Lowette, 2001).

EXAMPLES FOR CONTEXT-ADAPTIVE MOBILE SERVICES

Scholars focusing on mobile systems were among the first to publish research on adaptive systems. Olivetti and XEROX, two of the leading producers of copying machines at that time, were pioneers in this new research area of context-adaptive computing. In the early 1990s, their research facilities introduced the first prototypical adaptive systems. These early applications include automatic call-forwarding systems that are based on where in the office building the person who receives the call is located (Want, Hopper, Falcao, & Gibbons, 1992; Wood, Richardson, Bennett, Harter, & Hopper, 1997), browser software that can be adapted to specific locations (Voelker & Bershad, 1994), as well as location-aware shopping assistants (Asthana, Cravatts, & Krzyzanowski, 1994). These groundbreaking applications are not so much derived from data communication that is supported by mobile radio technology, but from communication that is connected to ubiquitous or, rather, pervasive computing. They primarily use infrared or radio transponders (so-called active badges). Special room sensors are able to detect users, who carry these transponders with them at all times. In later years, scholars developed various kinds of adaptive systems: these new developments include tour guides (Bederson, 1995; Long, Kooper, Abowd, & Atkeson, 1996; Davies, Cheverst, Mitchell, & Friday, 1999), software assistants for conference participants (Dey, Futakawa, Salber, & Abowd, 1999), and field researchers (Pascoe, 1998).

In the late 1990s, researchers proposed adaptive service concepts for commercial, content-related mobile radio services. One of the first studies offered a solution to the problem of content recipients' partially bound attention: this study developed an adaptive screen for GSM terminals which is able to change the screen's font and brightness in accordance with the room conditions and the user's activity. As a reaction to the deregulation of the cell-based user locating services in GSM-networks for commercial purposes in 2001, scholars proposed a number of other options for location-specific mobile services (Lehner, 2004). These suggestions include gas station search services that take locating and vehicle data into consideration, and adaptive multimedia applications for cars that are able to switch between different reception options (e.g., GSM or DVB), depending on the specific reception quality (Herden, Rautenstrauch, Zwanziger, & Planck, 2004). In addition to these cell-based location services, researchers have also discussed adaptive services based on GPS (Diekmann & Gehrke, 2003). Furthermore, some authors suggest individualizing concepts that are attuned to the special features of mobile terminals. These concepts would be able to adapt contents to individual preferences. In this context, it has to be distinguished between those individualization concepts that carry out the adaptation on the aggregation level (or rather, on a mobile portal—Smyth & Cotter, 2003; Kaspar & Hagenhoff, 2004) and those that aim at the individual service (Anderson, Domingos, & Weld, 2001). For the most part, scholars have discussed individualization techniques that are based on explicit information from users and on limited potentialities. Currently, however, research is also trying to find ways to distribute contents to a multitude of different types of mobile devices, a development that has become necessary because of the growing diversity of mobile devices that are equipped with varying hardware and software. One way to achieve this is to use markup transformations that are based on schematic libraries for different devices and standards such as WML, XHTML, und HTML. In order to identify a mobile device, various standards that allow users and providers to exchange the respective configuration during each data communication process have been developed. The best-known standards include the "Composite Capabilities/Preference Profiles" (CC/PP), which was developed by W3C in 2004 or its earlier implementation—the "User Agent Profile" (U-AProf)—by the WAP Forum in 2003. After identifying the respective device configuration, it is possible to adapt the syntax, for example on the basis of the style sheet transformation language, XSLT.

CONCLUSION

Most of the current examples of context-adaptive systems represent isolated solutions that are based on a closed-world assumption. At this point, these systems have little commercial value. This lack of commercial relevance is basically rooted in two main problems that have to be overcome in the future. On the one hand, the development of adaptive systems is very cost intensive. In addition, currently existing solutions are usually based on proprietary data models, which keep them from interacting with different systems and do not allow them to add additional contexts. On the other hand, an adaptive system is dependent on a comparatively large set of personal information that has to be gathered and processed automatically. This causes user concerns about possible abuses of personal data and intrusions of privacy. So far, neither legal measures nor technical control instruments have been able to eliminate users' apprehensions of permanent surveillance by a "big brother."

REFERENCES

Anderson, C., Domingos, P., & Weld, D. (2001). *Personalizing Web sites for mobile users.* Retrieved May 31, 2005, from http://www.cs.washington.edu/ai/proteus/www10.pdf

Asthana, A., Cravatts, M., & Krzyzanowski, P. (1994). An indoor wireless system for personalized shopping assistance. *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications*, Santa Cruz, CA, (pp. 69-74).

Bederson, B. (1995). Audio augmented reality: A prototype automated tour guide. *Proceedings of the Conference on Human Factors and Computing Systems*, Denver, (pp. 210-211).

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research.* Dartmouth Computer Science Technical Report TR2000-381. Retrieved October 31, 2005, from http://www.cs.dartmouth.edu/~dfk/papers/chen:survey-tr.pdf

Chen, G., Li, M., & Kotz, D. (2004, August). Design and implementation of a large-scale context fusion network. *Proceedings of the 1st Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services,* Boston.

Davies, N., Cheverst, K., Mitchell, K., & Friday, A. (1999). Caches in the air: Disseminating tourist information in the GUIDE system. *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications,* New Orleans.

Dey, A., & Abowd, G. (2000, June). The context toolkit: Aiding the development of context-aware applications. *Proceedings of the Workshop on Software Engineering for Wearable and Pervasive Computing,* Limerick, Ireland, (pp. 434-441).

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Dey, A., Futakawa, M., Salber, D., & Abowd, G. (1999). The conference assistant: Combining context-awareness with wearable computing. *Proceedings of the 3rd International Symposium on Wearable Computers (ISWC '99)*, San Francisco, (pp. 21-28).

Diekmann, T., & Gehrke, N. (2003). Ein framework zur nutzung situationsabhängiger dienste. In K. Dittrich, W. König, A. Oberweis, K. Rannenberg, & W. Wahlster (Eds.), *Lecture notes in informatics, informatik 2003. Innovative anwendungen* (Vol. 1, pp. 217-221). Bonn: Gesellschaft fuer Informatik.

Herden, S., Rautenstrauch, C., Zwanziger, A., & Planck, M. (2004). Personal information guide. In K. Pousttchi & K. Turowski (Eds.), *Mobile Economy: Proceedings of the 4th Workshop on Mobile Commerce*, Augsburg, (pp. 86-102).

Hightower, J., Brumitt, B., & Borriello, G. (2002, June). The location stack: A layered model for ubiquitous computing. *Proceedings of the 4th IEEE Workshop on Mobile Computing Systems & Applications,* Callicoon, New York, (pp. 22-28).

Kaspar, C., & Hagenhoff, S. (2004). Individualization of a mobile news service—a simple approach. In S. Jönsson (Ed.), *Proceedings of the 7th SAM/IFSAM World Congress,* Gothenburg.

Lehner, F. (2003). *Mobile und drahtlose informationssysteme.* Berlin: Springer-Verlag.

Lehner, F. (2004). Lokalisierungstechniken und location based services. *WISU, 2,* 211-219.

Long, S., Kooper, R., Abowd, G., & Atkeson, C. (1996). Rapid prototyping of mobile context-aware applications: The Cyberguide case study. *Proceedings of the 2nd Annual International Conference on Mobile Computing and Networking* (pp. 97-107), White Plains, NY.

Pascoe, J. (1998). Adding generic contextual capabilities to wearable computers. *Proceedings of the 2nd International Symposium on Wearable Computers,* Pittsburgh, PA.

Russel, S., & Norvig, P. (2003). *Artificial intelligence, a modern approach.* NJ: Pearson Education.

Schilit, B., Adams, N., & Want, R. (1994, December). *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications,* pp. 85-90.

Smyth, B., & Cotter, P. (2003). Intelligent navigation for mobile Internet portals. *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco.

Van Laerhoven, K., Aidoo, K., & Lowette, S. (2001). Real-time analysis of data from many sensors with neural networks. *Proceedings of the 5th IEEE International Symposium on Wearable Computers 2001,* (p. 115).

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

Wood, K., Richardson, T., Bennett, F., Harter, A., & Hopper, A. (1997). Global tele-porting with Java: Toward ubiquitous personalized computing. *Computer, 30*(2), 53-59.

KEY TERMS

Agent: A software system that is able to perceive its surroundings as events that are relevant for the application and that, in accordance with previously defined filter rules, causes actions in accordance with its perceptions.

Context: Any piece of information that can be used to characterize the situation of a person, a place, or an object in a way that is significant for the interaction between user and application.

Context-Adaptive System: An application that changes its behavior in accordance with information about the respective situation.

Location-Based Service: A service that takes the location of the user into consideration. A user's location can therefore be detected by using network-locating or terminal-locating.

Network-Locating: Detects an end device's position on the basis of network information.

Sensor Fusion: Refers to the translation of values from different locating systems using different measures (e.g., distance, angular separation, or geometrical position) into standardized measures and the transfer of these standardized measures into a joint data model.

Smart Space: The logical space that is covered by a computing device equipped with radio or infrared sensors within its sensor coverage.

Terminal-Locating: A procedure that enables a specially designed end device to locate its own position.

Context-Aware Mobile Geographic Information Systems

Slobodanka Djordjevic-Kajan
University of Nis, Serbia

Dragan Stojanović
University of Nis, Serbia

Bratislav Predić
University of Nis, Serbia

INTRODUCTION

A new breed of computing devices is taking more and more ground in the highly dynamic market of computer hardware. We refer to smart phones and PocketPCs, which redefine typical usage procedures we are all familiar with in traditional, desktop information systems. Dimensions of this class of computing devices allow users to keep them at hand virtually at all times. This omnipresence allows development of applications that will truly bring to life the motto: "availability always and everywhere."

Hardware and software characteristics of the aforementioned devices require a somewhat modified approach when developing software for them. Not only technical characteristics should be considered in this process, but also a general set of functionalities such an application should provide. Equally important is the fact that the typical user will be on the move, and his attention will be divided between the application and events occurring in his environment. Fundamentally new and important input to mobile applications is constantly changing the user environment. The term that is used most frequently and describes the user environment is a context, and applications that are able to independently interpret a user's context and autonomously adapt to it are named context-aware applications.

Recent developments in wireless telecommunications, ubiquitous computing, and mobile computing devices allowed extension of geographic information system (GIS) concepts into the field. Contemporary mobile devices have traveled a long way from simple mobile phones or digital calendars and phonebooks to powerful handheld computers capable of performing a majority of tasks, until recently reserved only for desktop computers. Advancements in wireless telecommunications, packet data transfer in cellular networks, and wireless LAN standards are only some of technological advancements GIS is profiting from. This mobile and ubiquitous computing environment is perfect incubation grounds for a new breed of GIS applications, mobile GIS. Advances in mobile positioning have given a rise to a new class of mobile GIS applications called location-based services (LBS). Such services deliver geographic information and geo-processing services to the mobile/stationary users, taking into account their current location and references, or locations of the stationary/mobile objects of their interests.

But the location of the user and the time of day of the application's usage are not the only information that shapes the features and functionalities of a mobile GIS application (Hinze & Voisard, 2003). Like other mobile and ubiquitous applications, mobile GIS completely relies on context in which the application is running and used. The full potential of mobile GIS applications is demonstrated when used in the geographic environment they represent (Raento, Oulasvirta, Petit, & Toivonen, 2005). Thus, development of mobile GIS applications requires thorough analysis of requirements and limitations specific to the mobile environment and devices. Practices applied to traditional GISs are usually not directly applicable to mobile GIS applications. Limitations shaping future mobile applications, including mobile GISs, are ranging from hardware limitations of client devices to physical and logical environment of the running application. Considering the fact that mobile applications are used in open space and in various situations, the ability of the application to autonomously adapt itself to a user's location and generally a user's context significantly increases the application's usability. Regardless of the type of LBS and mobile GIS application, the part of the system that is handling context is fairly independent and can be separately developed and reused. The proper management of contextual data and reasoning about it to shape the characteristics and functionalities of mobile GIS applications leads to a full context-aware mobile GIS.

The second section presents concepts of mobile GISs and context awareness, and the principles of how context can be incorporated into traditional GIS features adapted to mobile devices. Data structures and algorithms supporting context awareness are also given. The third section presents GinisMobile, a mobile GIS and LBS application framework

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

developed at Computer Graphics and GIS Lab, University of Nis, which demonstrates the concepts proposed in this article. The last section concludes the article, and outlines future research and development directions.

CONTEXT AWARENESS IN MOBILE GIS

Even though the concept of mobile GIS is in its infancy, technologies that were prerequisite for development of this niche of GIS applications are today widely available and well known to GIS developers. It is reasonable to expect that there are prototypes available demonstrating all the advantages mobile GIS offers to field fork personnel. ESRI, as one of the leading companies in the GIS field in its palette of products, offers a mobile GIS solution targeting the PocketPC platform. It is called ArcPad (http://www.esri.com/software/arcgis/ bout/arcpad.html). It is a general type of mobile GIS solution with open architecture allowing easy customization and tailoring according to a specific customer's needs. It therefore offers a set of basic GIS functionalities and tools that are used to extend application with functionalities needed for specific usage scenarios. ESRI bases its ArcPad on four basic technologies: mobile computing device (PocketPC), basic set of spatial analysis and manipulation tools, global positioning system (GPS), and wireless network communication interface.

Basic GIS functionality understandably supported by ArcPad is geographic maps visualization in the form of raster images. In order to avoid the need for maps conversion into some highly specialized proprietary raster map format, ArcPad supports usage of all of today's widely used raster image formats, like JPEG, JPEG 2000, and BMP, as well as MrSID, which is common in GIS applications. Thematically different maps in the form of raster images can be grouped into layers. Apart from raster type, layers can also contain vector data. Also, standard vector type data formats are supported, most importantly the shapefile format. That is the most common vector data format in use in GIS today and is also well supported by other ESRI GIS software like ArcInfo, ArcEditor, ArcView, ArcIMS, and others. Other optimizations which enable sufficient speed in handling spatial data include spatial indexing schemes. Spatial indexing significantly increases speed of spatial objects visualization and search, especially on portable devices with limited processing power. Indexes are prepared on other desktop-type ESRI applications, and afterwards are transferred to a mobile device and used by ArcPad. In order to support usage of ArcPad throughout the world, a majority of map projections are included.

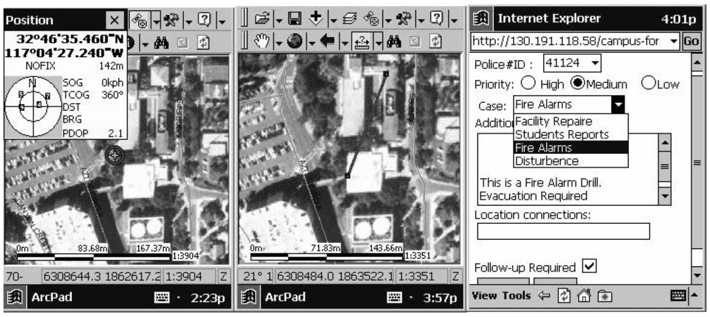
ArcPad is conceived as an integral part of the ESRI GIS platform consisting of other products, so there is the possibility of ArcPad functioning as a client for ArcIMS or Geography Network (http://www.geographynetwork.com/). Data is transferred to ArcPad using TCP/IP protocol and any sort of packet-based wireless networking technology (wireless LAN, GSM, GPRS, EDGE, 3G, etc.). Possibly the strongest advantage of ArcPad is its extensibility and adaptability. Forms used for thematic data input and manipulation are created and customizes independently using ArcPad Studio and Application builder development tools. Application toolbars can be adapted to specific user needs. More importantly, specific interfaces can be developed and added to ArcPad, enabling it to acquire data from different database types and sensors (GPS location devices, laser rangefinders, magnetic orientation sensors, etc.).

One academic project that encompasses the development of mobile GIS is "Integrated Mobile GIS and Wireless Image Servers for Environmental Modeling and Management," developed at San Diego State University (2002). The project includes an integrated GIS platform where, in the field, data collection must be performed using a mobile GIS client platform. Effectiveness of the developed system is tested in three different services: campus security, national park preservation service, and sports events. The development group's decision was not to develop a mobile GIS solution from scratch, but to upgrade and customize ArcPad. Similarly to other mobile GIS solutions, this project is based on modified client/server architecture. Fieldwork personnel are using a PocketPC device with a customized ArcPad version installed. Customization includes components developed specifically for testing on campus. PocketPC is connected with an external GPS device, and therefore it has constant access to user location information. Considering wireless communications, campus grounds are covered with a wireless LAN, and all client PocketPCs are equipped with WLAN adapters. The server side of this system includes a typical set of servers and tools from ESRI including ArcIMS and ArcGIS.

When this system is employed by the campus security service, field units use mobile GIS components to locate a reported incident location more easily and swiftly. Mobile GIS is also used to report new incidents to central. Following report-in, information about a new event taking place is momentarily available to all units. Therefore, reaction time is shortened and all patrolling units within campus are synchronized more easily.

Demonstration use case shows the field unit receiving a warning about a fire reported at the specified site. The closest field unit is being notified. Using the campus WLAN, the central ArcIMS server is contacted and a map of that part of the campus is acquired, as well as blueprints of buildings endangered by fire. The central server also contains thematic data about the estimated number of people in these buildings, evacuation plans, and similar information. Simultaneously, units on site can update fire reports with more detailed information and therefore shorten response time of other units enroute. The ArcPad application customized for this use and being used in this scenario is shown in Figure 1.

Figure 1. Mobile GIS implemented in San Diego State University campus security



Police#ID : 41124

Priority: ○ High ● Medium ○ Low

Case: Fire Alarms

This is a Fire Alarm Drill.
Evacuation Required

Follow-up Required ☑

Besides the location of the user, contemporary mobile GIS applications, such as the one previously described, lack the support for context awareness. Such support must be developed and integrated into the basic framework or platform on top of which the mobile GIS application is developed. But first we must define the context and basic principles of context awareness.

In interpersonal communication, a significant amount of information is transmitted without explicit communication of such information. If we take verbal communication as an example, nonverbal signs will significantly influence the completeness of verbally communicated data. We are referring to facial expressions, body posture, voice tone, and nearby objects and persons included in the past history of communications. All this is helping the process of interpretation of verbally transmitted data. In a typical human-machine communication, there is very little context information available in a form that can be interpreted by machine. Therefore our first step should be to define the context. No matter how obvious this may seem, the definition of the context influences significantly all the decisions in the further process of context-aware application development. Dey and Abowd (2000) give a relatively abstract definition of context influenced by their work on "context toolkit" architecture:

We define context as any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or physical or computational object.

Schilit and Theimer (1994) give a very concrete context definition which is therefore rather local in its application:

Context refers to location, identity of spatially nearby individuals and objects and changes that are relevant to aforementioned individuals and objects.

Summarizing numerous definitions of context, we can notice three aspects of context that are standing out:

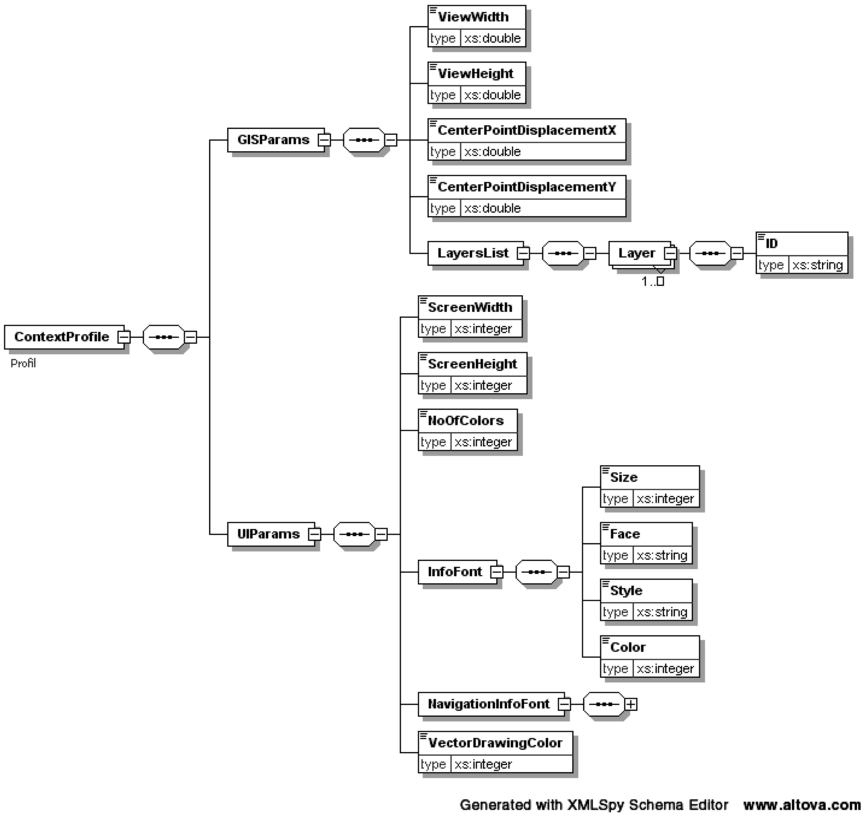
- **Technical Characteristics of the Environment:** Hereby we are mainly referring to technical characteristics of the client device, processing power, available memory capacity, display characteristics, as well as characteristics of network connections available to the device (bandwidth, latency, price, etc.).
- **Logical Characteristics of the User's Environment:** This group contains geographic location, identity of individuals and objects nearby, and general social situation.
- **Physical Characteristics of the User's Environment:** This group contains levels of noise, light, and movement parameters (speed, direction, etc.).

In the process of context modeling and management, the system can use information that is both automatically collected or manually entered by the user. Although the first approach is attractive and seems to be the only true manner of handling contextual data, we believe that manual input should not be excluded. Also, some characteristics of the context (e.g., user preferences, history, and predictions of actions) are much more easily acquired by manual input at the current level of advancements in context management algorithms.

The important step in development of a context-aware LBS and mobile GIS is to define the set of functionalities the application should provide to the user, implicitly or explicitly. Numerous types of contextual information produce adequately numerous potential functionalities. We can group them as follows:

- **Display of Information and Services:** In order to reduce user workload, the system adjusts the set of offered information and functions according to detected and deduced environment of the user. For a typical mobile GIS, a section of the map surrounding the current user's location is displayed. According to the user's speed and heading, the central point of the map view is chosen and the speed vector displayed. Also, font and color scheme are adjusted to the situ-

Figure 2. XML scheme describing profile


ContextProfile
Profil
GISParams
ViewWidth
type xs:double
ViewHeight
type xs:double
CenterPointDisplacementX
type xs:double
CenterPointDisplacementY
type xs:double
LayersList
Layer
1..∞
ID
type xs:string
UIParams
ScreenWidth
type xs:integer
ScreenHeight
type xs:integer
NoOfColors
type xs:integer
InfoFont
Size
type xs:integer
Face
type xs:string
Style
type xs:string
Color
type xs:integer
NavigationInfoFont
VectorDrawingColor
type xs:integer
Generated with XMLSpy Schema Editor www.altova.com

ation the user is in (e.g., the user is steering a vehicle at night).

- **Automated Execution of Commands:** An example would be a navigation GIS application that detects the user has missed the intersection and automatically initiates rerouting to find the new shortest path to destination.
- **Storage of Contextual Information:** Potential use of stored contextual information would be to enable application to autonomously extract user preferences from previous actions using data mining techniques.

The user context that is of interest in LBS and mobile GIS applications is classified into specific classes. Each class of contextual information is assigned a context variable. Often, in other papers published by researchers in this field, authors have noticed a hierarchical structure of context information, so some sort of graph structure is used for context representation (Meissen, Pfennigschmidt, Voisard, & Wahnfried, 2004). Since one class contains contextual information of various levels of generality, the most appropriate data structure for representing contextual information is directed acyclic graph. This data structure is the closest match to human cognition of structure and connections existing within a context data class. Another advantage of the hierarchical context model is the possibility to narrow down the choice of possible actions induced by a detected context. In this manner, a set of rules used by the rule-based expert system is kept to a minimum of candidate rules (Biegel & Cahill, 2004). The rule-based expert system is used to perform generalization of raw contextual data acquired from sensors and contained in leaf nodes. In this manner a "vertical" structure within each context class is built. As an example of a rule-based expert system that is widely used in literature, we have opted for the C Language Integrated Production System (CLIPS, 2006). The main advantage of CLIPS in our case is the existence of jCLIPS, a library that enables Java programs to use the CLIPS engine, embedding it in a Java code (jCLIPS, 2005).

The typical context data flow path in a context-aware application is as follows: raw data is collected by connected sensors, and the software interface associated with each of the sensors converts the data into facts and stores the facts into the expert system. After each modification of a rule set, CLIPS executes a generalization process and generates the new facts at higher levels of generality. Also, the possibility of performing action is tested. The action is represented by forming an XML file containing configuration parameters for the client device. This XML file generally describes a profile that a context-aware application will use as a response to a context change. The XML scheme of such a profile is shown in Figure 2.

Figure 3. The XML file representing the user profile according to the user's context

```
<?xml version="1.0" encoding="UTF-8"?>
<ContextProfile xmlns="http://gislab.elfak.ni.ac.yu/bpredic" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://gislab.elfak.ni.ac.yu/bpredic:\Dragan\CONTEXT\sema.xsd">
 <GISParams>
     <ViewWidth>1000</ViewWidth>
     <ViewHeight>700</ViewHeight>
     <CenterPointDisplacementX>850</CenterPointDisplacementX>
     <CenterPointDisplacementY>0</CenterPointDisplacementY>
     <LayersList>
             <Layer>
                     <ID>Gas_Stations</ID>
             </Layer>
             <Layer>
                     <ID>Fast_Food_Restaurants</ID>
             </Layer>
     </LayersList>
 </GISParams>
 <UIParams>
     <ScreenWidth>200</ScreenWidth>
     <ScreenHeight>320</ScreenHeight>
     <NoOfColors>4092</NoOfColors>
     <InfoFont>
             <Size>12</Size>
             <Face>Courier</Face>
             <Style>Normal</Style>
             <Color>Yellow</Color>
     </InfoFont>
     <NavigationInfoFont>
             <Size>24</Size>
             <Face>Arial</Face>
             <Style>Bold</Style>
             <Color>Red</Color>
     </NavigationInfoFont>
     <VectorDrawingColor>Blue</VectorDrawingColor>
 </UIParams>
</ContextProfile>
```

The particular profile transferred to the client is represented as an XML file described in Figure 3.

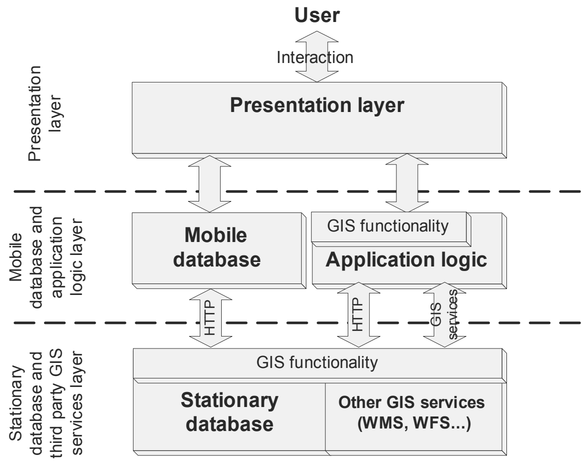
GinisMobile: CONTEXT-AWARE MOBILE GIS FRAMEWORK

To support efficient development of mobile GIS solutions, GinisMobile as a component mobile GIS framework is developed in the Laboratory for Computer Graphics and Geographic Information Systems (CG&GIS) Lab at the University of Nis. This framework represents the result of continuous development and advancement of GIS frameworks intended for rapid development of desktop and Web GIS applications (Ginis and GinisWeb). Well-tested GIS concepts of Ginis and GinisWeb have been supplemented with application components from the mobility domain (Predic & Stojanovic, 2004).

Mobile GIS architecture is somewhat similar to architectures used for Web or WAP GIS solutions (Fangxiong & Zhiyong, 2004). An extended client/server approach is used as the basis. The most frequent modification of this architecture, which is used in commercial solutions, is the three-tier model. It consists of the presentation layer, and the GIS logic layer including mobile database components and external GIS services layer. The third layer encompasses GIS services like Web map service (WMS) and Web feature service (WFS) (OGC, 2003). The main advantage of this approach is clear separation of functionalities into independent modules, easily upgradeable and substitutable. The grouping of layers in the case of mobile GIS is shown in Figure 4.

The presentation layer is tasked with information visualization (spatial data in the form of maps and attributes in alphanumerical tabular form). This layer also receives users' requests and queries, interprets them, and activates corresponding functions in the application logic layer. Using adapted HTTP protocol, the client is supplied with static

Figure 4. Layered architecture of mobile GIS



sections of maps in the form of raster images and vector type spatial data which is XML encoded. XML encoding also contains attribute data. All data the user interacts with and can change are classified and transferred in the vector form. All other data, regardless of their form when stored on the server side, are rasterized and transferred to the client in the form of raster images. Since the data's role is only auxiliary, this is the most appropriate form of visualization.

Database component must contain the functionality of partial replication of the data subset that is of interest to the individual user and data synchronization with local data storage located on the mobile device (Huang & Garcia-Molina, 2004). In our usage scenario (mobile GIS), the client possesses significant processing power and memory capacity which can therefore be used for performance improvement of this typical layered architecture. That is the reason the client-side component in the case of mobile GIS is usually named 'rich client' (Predic, Milosavljevic, & Rancic, 2005).

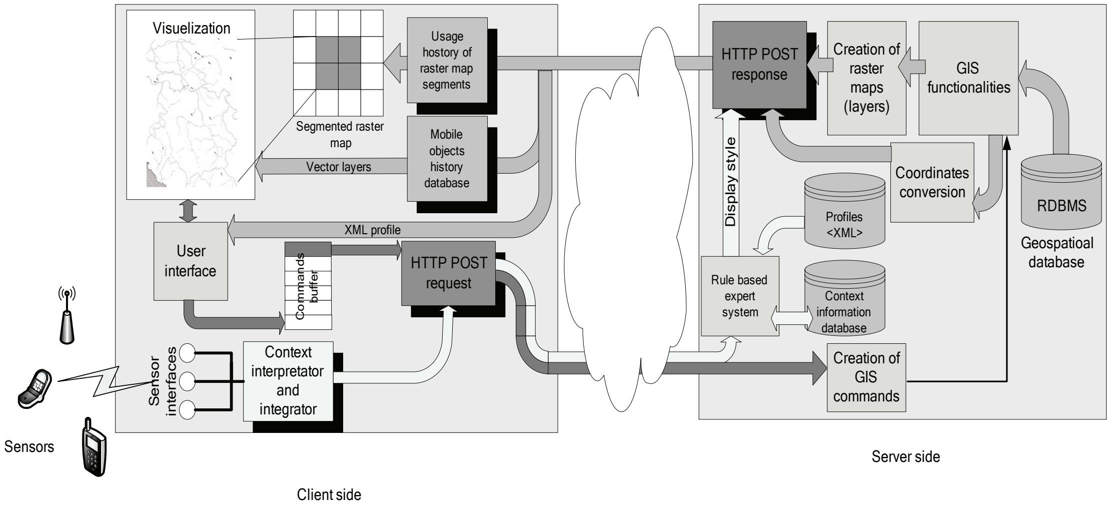
The mobile database and application logic layer is physically located at the client device according to its significant processing capacities. Application logic, which is also physically located at the client device, contains a portion of GIS functionalities, basic functionalities which can be performed on the client side solely without data transfer to/from the server side. The stationary database and third-party GIS services layer (e.g., Web map server) is physically located on the server side. The stationary database contains the complete set of data available. The mobile GIS client is supplied only with a subset of available data, a subset that is of direct interest to the individual client (fieldwork team). Determining the scope and volume of this subset is the task of the GIS functionalities component located in this layer. Other GIS services belonging to this layer (WMS, WFS) are also controlled by GIS functionalities of this layer. WMS is tasked with supplying raster map segments which are used at the client to form a continuous geo-referenced map. WFS provides data about geographic features in the vector format encoded in Geography Markup Language (GML, 2004).

To support development of context-aware LBS and mobile GIS applications, we have developed and integrated context-aware support and components into GinisMobile, an LBS and mobile GIS application framework (Predic et al., 2005). GinisMobile is a mobile extension of GinisWeb, a Web GIS application framework (Predic & Stojanovic, 2004). As such it includes support for management and presentation of raster and vector spatial data, as well as dynamic data about mobile objects (Stojanovic, Djordjevic-Kajan, & Predic, 2005). The first obstacle encountered when developing LBS and mobile GIS applications is a highly constrained mobile client platform. Therefore, these applications are already aware of the hardware characteristics of the device it is running on and able to automatically adapt to it, enabling full utilization of the device's capabilities. The type of sensors relevant to context-aware LBS and mobile GIS applications are widely available, and either are already integrated in modern devices (GPS sensors, Bluetooth radios, level of light sensors, etc.) or are available as add-on devices connected via PAN (Bluetooth). Each of these sensors implements its internal data format. Therefore, each sensor has a software interface attached to it. Its task is to convert data from the format used internally by the sensor into a format appropriate to the application. Contextual data concerning technical characteristics of the device are accessible directly by the application and therefore do not require a separate software interface. A compiled set of contextual data is encoded according to a defined XML scheme and transferred to the server for analysis and storage. The proposed architecture of GinisMobile, a context-aware LBS and mobile GIS framework, is illustrated in Figure 5.

This model requires minimal changes to starting a mobile GIS architecture and minimizes processing requirements on the client side. On the server side, context information is handled separately from the user commands. It is inserted into the rules and facts database, and is analyzed by a rule-based expert system. Every change in the rules and facts database can lead to either insertion of new context information at the higher logical level into the database or entering a new state. Reaching a new state results in picking out a profile that most adequately fits new change in the user's context. Handling of other spatial data is the same as in Predic and Stojanovic (2005) and will not be further discussed in this article. Profile, packed with static spatial layers (rasterized to a single map layer) and a dynamic map component (e.g., moving objects), is transferred to the client.

On the client side, the XML profile is parsed and used to customize the user interface. Rasterized layers are stored on internal cache and displayed. Static objects are presented on the background map according to display settings, profile,

Figure 5. Mobile GIS architecture with context-aware support


Visuelization
Segmented raster map
Vector layers
Usage hostory of raster map segments
Mobile objects history database
XML profile
User interface
Commands buffer
HTTP POST request
Sensor interfaces
Context interpretator and integrator
Sensors
Client side
HTTP POST response
Creation of raster maps (layers)
GIS functionalities
Display style
Coordinates conversion
Profiles <XML>
RDBMS
Geospatioal database
Rule based expert system
Context information database
Creation of GIS commands
Server side

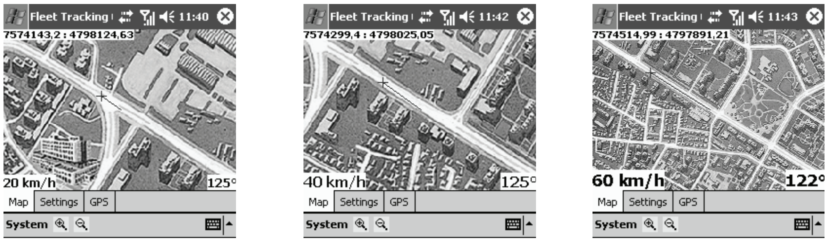
and with appropriate symbols. Finally, moving objects are superimposed on the map display. Since raster segments are static in nature and change rarely, we keep a local cache of frequently used segments. This approach speeds up the visualization process significantly. The adopted least recently used (LRU) algorithm is used to keep memory requirements minimal.

To test the context awareness support and context-aware components built into the GinisMobile framework, we have developed a mobile GIS application for vehicle navigation and fleet tracking on top of the GinisMobile. The application setting assumes that the sensors connected to the user's device are able to determine speed and direction, time of a day, and levels of noise and light. The higher level of contextual information is deduced based on basic contextual data. Based on deduced facts and rules within the knowledge-based engine, the server recognizes that the user drives a vehicle during the evening/night hours. According to this information, an appropriate profile is constructed which describes the user interface with night colors. The navigation data are displayed on the screen with appropriate font size according to the speed of the user's vehicle. Also, appropriate zoom level is chosen with the user's location displaced from the view center. In this manner the user is enabled to see more of the map in front of him. Finally, the view contains vector speed as a reference.

We assume the existence of a GPS receiver attached to the mobile device since this is a very common type of sensor today. It provides data on a user's geographic location as well as motion data (speed and direction). Another "sensor" relies on time of day to detect light conditions (day/night). More specifically, a level-of-light sensor that is present on mobile devices available on the market could be used for this purpose for additional accuracy.

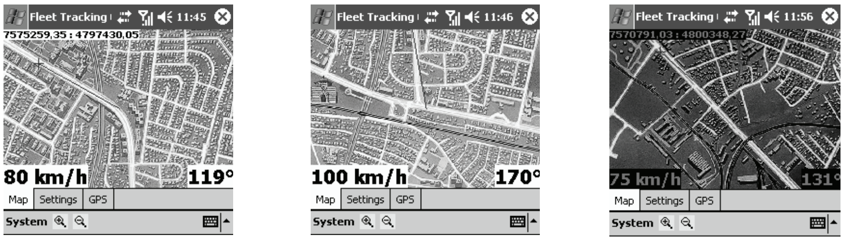
The role of context detection and interpretation subsystem is to decrease the workload needed to operate a typical vehicle navigation system and therefore increase safety. In the demonstration application the following scenario is employed: a mobile user drives a vehicle in the urban environment during the day and at night. Screenshots are taken at different levels of adaptation that an LBS application has performed autonomously in response to changing user context. Figure 6(a) shows a user traveling at 20km/h

Figure 6. Screenshots taken from the sample mobile GIS application


Fleet Tracking
11:40
7574143,2 : 4798124,63
20 km/h
125°
Map
Settings
GPS
System
Fleet Tracking
11:42
7574299,4 : 4798025,05
40 km/h
125°
Map
Settings
GPS
System
Fleet Tracking
11:43
7574514,99 : 4797891,21
60 km/h
122°
Map
Settings
GPS
System

(a) (b) (c)

Figure 6. Continued



(d) *(e)* *(f)*

along a city street. It is worth noticing that the user's location (indicated by the cross symbol) is decentralized and a velocity vector is drawn on the map view. The map view also includes speed and heading.

As the user increases his speed, the font size for displayed motion data (speed and heading) is increased, the amount of map view decentralization is increased, and the velocity vector is updated accordingly. This is illustrated in Figure 6(b). As the speed further increases above a certain threshold (Figure 6(c)), the map view zoom scale is changed (decreased). This, along with additional decentralization of a map view, allows the user to see more of the map laying in front of him, in the direction of the velocity vector. The effects of further increase of speed and change in direction of velocity vector are shown in Figures 6(d) and 6(e). When the application detects night conditions, it switches to using a set of colors customized to night conditions. The map view customized to night conditions is shown in Figure 6(f). This choice of colors minimizes the distraction effect for the user (driver).

CONCLUSION

Considering the general trend in development of information technologies and the computer industry in general, we can notice a constant migration into mobile and ubiquitous computing (Hinze & Voisard, 2003). With further development of wireless communication technologies, data stored in heterogeneous distributed databases will be available at any specific instance in time and at any location. The most important beneficiaries of this newly introduced concept of mobile GIS will be professional users whose job descriptions include a lot of field work with spatial data. Considering the current state of the art, we believe that mobile GIS applications are perfect testing grounds for the context-aware concept. Being used in unconstrained free space, context awareness considerably enhances usability of the mobile GIS applications. Hereby, context-aware applications are a super set of location-based services. As this article has stressed, location information is only one class, although very frequently used, of context information. This information will be used to automate many procedures and decisions, and relieve the user of the repeatable and tedious tasks of frequent reconfiguration.

REFERENCES

Biegel, G., & Cahill, V. (2004, March 14-17). A framework for developing mobile, context-aware applications. *Proceedings of the 2nd IEEE Conference on Pervasive Computing and Communications (Percom 2004)*, Orlando, FL.

CLIPS. (2006). *Version 6.24: A tool for building expert systems.* Retrieved from http://www.ghg.net/clips/CLIPS.html

Dey, A. K., & Abowd, G. D. (2000, April 1-6). Towards a better understanding of context and context-awareness. *Proceedings of the Workshop on the What, Who, Where, When and How of Context-Awareness* (affiliated with CHI 2000), The Hague, The Netherlands.

Fangxiong, W., & Zhiyong, J. (2004, July 12-23). Research on a distributed architecture of mobile GIS based on WAP. *Proceedings of the ISPRS Congress,* Istanbul, Turkey.

GML. (2004). *Geography Markup Language (version 3.1.1)—encoding specification.* Retrieved from http://portal.opengeospatial.org/files/?artifact_id=4700

Hinze, A., & Voisard, A. (2003, July 24-27). Location- and time-based information delivery. *Proceedings of the 8th International Symposium on Spatial and Temporal Databases,* Santorini Island, Greece.

Huang, Y., & Garcia-Molina, H. (2004). Publish/subscribe in a mobile environment. *Wireless Networks, 10*(6), 643-652.

JCLIPS. (2005). Retrieved from http://www.cs.vu.nl/~mrmenken/jclips/#developer

Meissen, U., Pfennigschmidt, S., Voisard, A., & Wahnfried, T. (2004). Context and situation awareness in information logistics. *Proceedings of the Workshop on Pervasive Information Management* (held in conjunction with EDBT 2004).

Open GIS® Reference Model. (2003). *Version 0.1.2, Open GIS Consortium, Reference Number: OGC 03-040.* Retrieved from http://www.opengis.org/specs/?page=orm

Predic, B., & Stojanovic, D. (2004, March 8-12). XML integrating location based services with Web based GIS. *Proceedings of YU INFO 2004,* Kopaonik, Serbia, and Montnegro.

Predic, B., & Stojanovic, D. (2005, May 26-28). Framework for handling mobile objects in location based services. *Proceedings of the 8th Conference on Geographic Information Science (AGILE 2005)* (pp. 419-427), Estoril, Lisbon, Portugal.

Predic, B., Milosavljevic, A., & Rancic, D. (2005, June 5-10). RICH J2ME GIS client for mobile objects tracking. *Proceedings of the XLIX ETRAN Conference,* Budva.

Raento, M., Oulasvirta, A., Petit, R., & Toivonen, H. (2005). Context phone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing—Mobile and Ubiquitous Systems,* (April-June), 51-59.

San Diego State University. (2002). *Integrated mobile GIS and wireless Internet image servers for environmental monitoring and management.* Retrieved from http://map.sdsu.edu/mobilegis/photo_mtrp.htm

Schilit, B., & Theimer, M. (1994). Disseminating active map information to mobile hosts. *IEEE Network, 8*(5), 22-32.

Stojanovic, D., Djordjevic-Kajan, S., & Predic, B. (2005, December 15-16). Incremental evaluation of continuous range queries over objects moving on known network paths. *Proceedings of the 5th International Workshop on Web and Wireless Geographical Information Systems* (LNCS 3833, pp. 168-182), Lausanne, Switzerland. Berlin: Springer-Verlag.

KEY TERMS

Context-Aware Application: Application that posses the ability to autonomously and independently detect and interpret a user's environment parameters, and adapts its performance and functionalities according to detected context.

Geographic Information System (GIS): Information system that stores, analyzes, and presents data about geographic entities.

Geography Markup Language (GML): The XML grammar defined by the Open Geospatial Consortium (OGC) to express geographical features.

Global Positioning System (GPS): Satellite-based system that is using radio triangulation techniques to determine geographic position time and speed of user.

Location-Based Service (LBS): Service that delivers geographic information and geo-processing services to a mobile/stationary user, taking into account the user's current location and references, or locations of the stationary/mobile objects of the user's interests.

Web Feature Service (WFS): Web-accessible service that provides data about geographic entities encoded in GML using HTTP protocol.

Web Map Service (WMS): Web-accessible service that provides geo-referenced raster map data using HTTP protocol.

Context-Aware Systems

Chin Chin Wong
British Telecommunications (Asian Research Center), Malaysia

Simon Hoh
British Telecommunications (Asian Research Center), Malaysia

INTRODUCTION

Fixed mobile convergence is presently one of the crucial strategic issues in the telecommunications industry. It is about connecting the mobile phone network with the fixed-line infrastructure. With the convergence between the mobile and fixed-line networks, telecommunications operators can offer services to users irrespective of their location, access technology, or terminal.

The development of hybrid mobile devices is bringing significant impact on the next generation of mobile services that can be rolled out by mobile operators. One of the visions for the future of telecommunication is for conventional services such as voice call to be integrated with data services like e-mail, Web, and instant messaging. As all these new technologies evolve, more and more efforts will be made to integrate new devices and services. New markets for services and devices will be created in this converged environment. Services become personalized when they are tailored to the context and adapted to changing situation.

A context-aware network system is designed to allow for customization and application creation, while at the same time ensuring that application operation is compatible not just with the preferences of the individual user, but with the expressed preferences of the enterprise or those which own the networks. In a converged world, an extended personalization concept is required. The aspects covered include user preferences, location, time, network, and terminal; these must be integrated and the relationships between these aspects must be taken into consideration to design business models. Next-generation handsets are capable of a combination of services available on a personal digital assistant (PDA), mobile phone, radio, television, and even remote control. This kind of information and communications technology and mobile services together form one of the most promising business fields in the near future.

The voice average revenue per user (ARPU) is declining, the competition is getting fiercer, and voice over Internet protocol (VoIP) is entering the market with aggressive pricing strategies. Fixed mobile convergence should help in this context by providing converged services to both consumer and small-business users. For telecommunication companies it is now crucial to attempt to identify concrete applications and services for commercial offerings based on fixed mobile convergence which go beyond the current hype. Market scenarios and business models for such fixed mobile convergence solutions will be required and are therefore valuable for future strategy decisions.

This article examines market aspects, user requirements, and usage scenarios to come up with a roadmap and suggestions on how to deal with this matter.

CURRENT AND FUTURE TRENDS

In the past, user movement has often implied interruption of service. With the advent of pocket-size computers and wireless communication, services can be accessed without interruption while the entity using the services is moving (Floch, Hallsteinsen, Lie, & Myrhaug, 2001). There is a strong need for seamless access. Convergence has been taking place for years now. A study performed by the European Commission (1997) defines convergence as allowing both traditional and new communication services, whether voice data, sound, or pictures to be provided over many different networks. An excellent example of convergence in the telecommunications industry is the IP multimedia subsystem (IMS).

Similar to other emerging industries, fixed mobile convergence is characterized by a continuously changing and complex environment, which creates uncertainties at technology, demand, and strategy levels (Porter, 1980). Porter (1980) asserts that it is possible to generalize about processes that drive industry evolution, even though their speed and direction vary. According to Ollila, Kronzell, Bakos, and Weisner (2003), these processes are of different types and are related to:

- market behavior;
- industry innovation;
- cost changes;
- uncertainty reduction; and
- external forces, such as government policy and structural change in adjacent industries.

Each evolutionary process recognizes strategic key issues for the companies within the industry, and their effects

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

are usually illustrated as either positive or negative from an industry development viewpoint. For example, uncertainty reduction is an evolutionary process that leads to an increased diffusion of successful strategies among companies and the entry of new types of companies into the industry. Both of these effects are believed to contribute to industry development with regards to the fixed mobile convergence value Web.

The technological uncertainties are usually caused by fast technological development and the battles for establishing standards, which are common in the beginning stages of the lifecycle of a specific industry as a result of a technological innovation (Camponovo, 2002). Concerning demand, regardless of the generalized consensus about the huge potential of fixed mobile convergence, there are many uncertainties about what services will be developed, whether the users are ready to pay for them, and the level and timeframe of their adoption (Camponovo, 2002).

While the wireless industry is often cited as an example of a rapidly changing sector, the period from 2001-2005 could (in some respects) be regarded as relatively stable (Brydon, Heath, & Pow, 2006). Mobile operators have made the vast majority of their service revenue from simple voice telephony and text messaging, while their value chain has remained largely undisturbed (Brydon et al., 2006). However, new services, alternative technologies, and an evolving competitive landscape mean that the possibility of substantial industry change over the course of the next five to 10 years cannot be discounted (Brydon et al., 2006).

The telecommunication industry has experienced several waves of changes from the introduction of wired telephony to wireless telephony, and it is currently heading towards fixed-mobile convergence. Users become more demanding: a "user-centric" and not "network-centric" approach is needed.

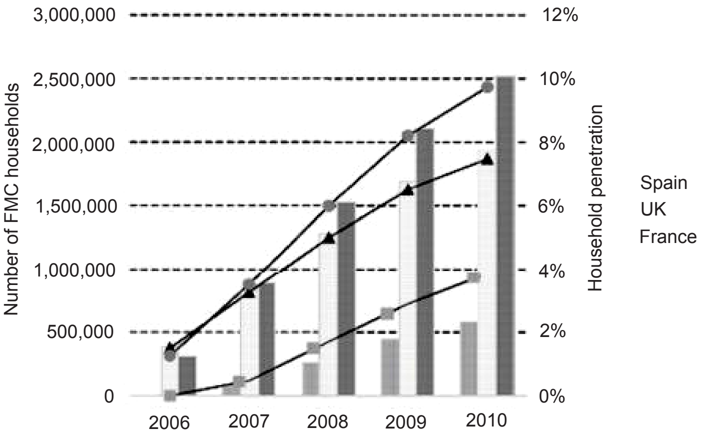
According to Hellwig (2006), many fixed operators lose their market dominance and merge units (fixed and mobile). New technologies and new actors (e.g., VoIP, Wi-Fi operators) coming into the picture are driving the adoption of fixed mobile convergence. The formation of new roles in the communication industry—including brokers, aggregators, alliances, and cooperation—have further pushed the stakeholders to take aggressive strategies to gain competitive advantage.

However, since new roles have been introduced, it is unclear how the market acceptance in the near future will be. Existing business models might not be applicable in the new business environment. The lack of terminal devices at the moment also hinders the diffusion.

In markets where there are high levels of fixed-mobile substitution and where broadband penetration and wireless local area network (WLAN) diffusion in the home are accelerating, it is most likely that consumers will be drawn to fixed mobile convergence, provided the cost savings and added convenience of carrying one device are apparent to the consumer (McQuire, 2005). According to the Yankee Group (2005), almost one-third of users make more calls within the home using their mobile phone than their landline. The trend is stronger among younger respondents. Figure 1 shows the number of fixed mobile convergence households in Spain, the UK, and France from 2006 to 2010.

The increasing need for a personal communication device that can connect to any type of network—a mobile network, IP network, or even public switched telephone network (PSTN)—and that supports all voice and text-based communication services drives the development of context-aware systems. The primary objective of the system is to facilitate acquisition, translation, and representation of context information in a structured and extensible form, in order to enable the development and enhancement of functionality of network

Figure 1. Number of fixed mobile convergence households in Spain, the UK, and France 2006-2010 (Hellwig, 2006)



resources, personalized according to each individual's needs. The secondary goal is to facilitate rapid development and deployment of services and applications through a defined framework, which can maintain interoperability between different services and domains.

An example of a context-aware system would be BT's proposed Context Aware Service Platform (CASP). In June 2004, NTT DoCoMo, BT, and a number of other incumbent operators from around the world formed the Fixed Mobile Convergence Alliance (2006), with the objective of developing common technology standards and low-cost devices for integrated fixed-mobile services. The CASP middleware mentioned is the interpretation and one of BT's visions for the development of a converged platform. The salient features of the proposed product include:

- **User-Centered Operability:** One important requirement for a heterogeneous network environment is the ability to instantaneously optimize services for individual users without the need for them to perform any annoying operations. CASP aims to provide transparent connectivity between users with devices and surrounding communication resources. It is able to recognize users' situations and environmental information automatically.
- **Ease of Service Provisioning:** The proposed platform and generic framework guidelines in respect to security, data integrity, non-repudiation, registration, subscription, and quality of service (QoS) for all services will be made available. It offers standard interfaces for all services which enable easier access to a less complex network, with common operation and management, maintenance and training, as well as a common environment for services development and delivery.
- **Interoperability of Shared Services:** The proposed platform provides a common specification for services to guarantee the interoperability between shared services in the communication networks. Specific context information with respect to specific aspects characterizing a service or entity can be expressed in an eXtensible Markup Language (XML)-based instance document.
- **Unified Identity:** In a true seamless access communication world, every user or communication object is represented by a unified identity. A session initiation protocol (SIP) address (e.g., simon.hoh@bt.com) can be used to uniquely identify a user or communication object even when it moves across different networks or between different devices. By having identity management, it simplifies mobility management, security management, and unified user profile management.
- **Dynamic User Interface (UI) on Shared Device:** Through the proposed platform, the user can have a shared device that can connect and interact with the ubiquitous communication objects nearby. Each networked object or entity such as cameras, scanners, printers, video players, and so forth can be represented by a different UI based on its own dynamic profile and thus can react intelligently to events in the communication space.
- **Context-Enabled Adaptive Service:** The heterogeneity of the converged networks, in terms of network capacity and terminal capabilities, is expected to cause unpredictable changes of network condition. The traditional QoS mechanisms, which do not take the presence of mobility and seamless connectivity into consideration, are not sufficient to guarantee a stable service. Thus, the use of adaptive services being able to change their settings to adapt to the available network resources is a must. CASP enables dynamic selection of the settings used by multimedia services and applications during a multimedia session based on the context of the surrounding environments.

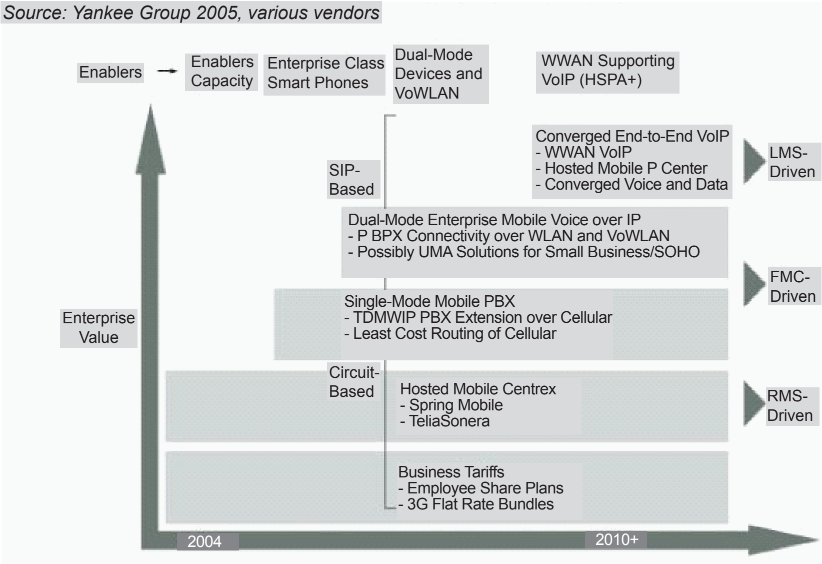
In the near future, stakeholders in the industry will move towards IP-based transport, call control, and service creation and delivery platform functionality. They will follow and adopt developments in the 3rd Generation Partnership Project (3GPP) (http://www.3gpp.org/), European Telecommunications Standards Institute (ETSI) (TISPAN) (http://www.etsi.org/), and Internet Engineering Task Force (http://www.ietf.org/) Next-Generation Network (http://www.ngni.org/) to support open interfaces and avoid interconnection and co-operation incompatibilities. In addition, the will: (1) support IP-based signaling and addressing, media negotiation, QoS, and security mechanisms; (2) support a very large variety of multimedia, banking, and mobile office applications seamlessly across different networks; and (3) adopt seamlessly to the network characteristics and device used. The players in the industry must also consider entering new positions in the value network by taking on new roles and working in cooperation with other telecommunication companies (Hellwig, 2006).

Figure 2 shows how and when future enterprise telephony services will embrace convergence.

USAGE SCENARIOS

Cheryl subscribes to an Internet VoIP service to save money on calls to her family and university friends who are now spread around the globe, but since her mobile operator utilizes unlicensed mobile access (http://www.umatechnology.org/) technology, she is now able to enjoy cheaper calls by using her mobile phone and connecting to her home WLAN or public hotspots. Her mobile device also enables a number of rich services that enable her to communicate with her friends via voice, video, as well as text.

Figure 2. Future enterprise telephony services will embrace convergence



Source: Yankee Group 2005, various vendors

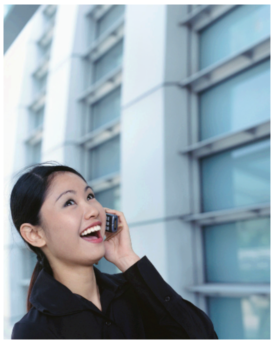
Enablers
Enablers Capacity
Enterprise Class Smart Phones
Dual-Mode Devices and VoWLAN
WWAN Supporting VoIP (HSPA+)
Converged End-to-End VoIP
- WWAN VoIP
- Hosted Mobile P Center
- Converged Voice and Data
LMS-Driven
SIP-Based
Dual-Mode Enterprise Mobile Voice over IP
- P BPX Connectivity over WLAN and VoWLAN
- Possibly UMA Solutions for Small Business/SOHO
FMC-Driven
Enterprise Value
Single-Mode Mobile PBX
- TDMWIP PBX Extension over Cellular
- Least Cost Routing of Cellular
Circuit-Based
Hosted Mobile Centrex
- Spring Mobile
- TeliaSonera
RMS-Driven
Business Tariffs
- Employee Share Plans
- 3G Flat Rate Bundles
2004
2010+

When she was on vacation in Hawaii recently, she was able to show the pictures she had taken with her mobile device to a colleague while he was talking to her over a VoIP call, and later sent a "wish you were here" video message to her parents.

Since all her communications are unified in a single device, Cheryl's friends and family can always reach her, either by voice, text, instant message, video call, or any other means, while Cheryl can use presence to broadcast her availability to her contacts (such as "in a meeting" or "traveling"), as well as manage the incoming communication depending on the context of what she is doing.

Now, she can keep her long-distance bills lower by using VoIP, but since it is in her mobile device, she does not have to be sitting in front of her personal computer (PC) to use it. And whereas some of the PC-based services Cheryl previously used were cumbersome to set up, and she had separate providers for her telecom services, she now gets all these functionalities in a bundled offering from a single operator, and all technology takes care of itself, working invisibly to her through a user-friendly interface.

Figure 3.



Cheryl is a recent university graduate whose work as a researcher means she has a busy travel schedule. She often travels to several places to attend international conferences and workshops. She uses a multi-radio mobile device and a mobile VoIP service subscription to keep in touch with her friends and family whether she is at home or on the move.

CONCLUSION

Context-aware systems, when made available to the end users, will be greatly valued by them. Consumers will be in a more interactive environment that could help them to take care of small yet related matters automatically. Any possible devices around them could be used to bridge any services offered, giving them the familiarity they preferred. Users would always have the option to alter the service execution or switch it off anytime they like.

Meanwhile from the network perspective, knowing the situation of the network and each network node's role could enable an adaptive and intelligent network. The capabilities such as self-healing, autonomous utilization optimization, and self-reconfiguration to adapt for changes could also be enabled with context-sensitive service logic.

The CASP provides a stable and robust environment for the context-aware service developers and operators. This stable environment is extremely important for them to have the accurate anticipated outcome and have the flexibility on changes. On top of the stable environment, the platform will be assisting the service execution to reduce the complexity for the service creation.

REFERENCES

Brydon, A., Heath, M., & Pow, R. (2006). *Scenarios for the evolution of the wireless industry in Europe to 2010 and beyond.* Analysys Research.

Camponovo, G. (2002). *Mobile commerce business models.* Paper presented at the International Workshop on Business Models, Lausanne, Switzerland.

European Commission. (1997). *Towards an information society approach.* Green paper on the convergence of the telecommunications, media and information technology sectors, and the implications for regulation. European Union.

Floch, J., Hallsteinsen, S., Lie, A., & Myrhaug, H. I. (2001, November 26-28). *A reference model for context-aware mobile services.* Tromsø, Norway: Norsk Informatikkonferanse.

FMCA. (2006). Retrieved from http://www.thefmca.com/

Hellwig, C. (2006). *New business and services by converging fixed and mobile technologies and applications.* T-Systems.

McQuire, N. (2005). *Residential fixed-mobile convergence heats up but SME Is next frontier.* Yankee Group.

Ollila, M., Kronzell, M., Bakos, M., & Weisner, F. (2003). *Mobile entertainment industry and culture: Barriers and drivers.* UK: MGAIN.

Porter, M. (1980). *Competitive strategy.* New York: The Free Press.

Yankee Group. (2005). *European mobile user survey.* Yankee Group.

KEY TERMS

eXtensible Markup Language (XML): A specification developed by the W3C, XML is a pared-down version of the Standard Generalized Markup Language (SGML), designed especially for Web documents. It allows designers to create their own customized tags, enabling the definition, transmission validation, and interpretation of data between applications and between organizations.

Internet Protocol Multimedia Subsystem (IMS): A standardized next-generation networking (NGN) architecture for telecommunication companies that want to provide mobile and fixed multimedia services. It uses a VoIP implementation based on a 3GPP standardized implementation of session initialization protocol (SIP), and runs over the standard Internet protocol (IP). Existing phone systems (both packet-switched and circuit-switched) are supported.

Public-Switched Telephone Network (PSTN): The international telephone system based on copper wires carrying analog voice data. This is in contrast to newer telephone networks based on digital technologies, such as the integrated services digital network (ISDN) and fiber distributed data interface (FDDI).

Quality of Service (QoS): A networking term that specifies a guaranteed throughput level. One of the biggest advantages of asynchronous transfer mode (ATM) over competing technologies such as frame relay and fast ethernet is that it supports QoS levels. This allows ATM providers to guarantee to their customers that end-to-end latency will not exceed a specified level.

Session Initiation Protocol (SIP): An application-layer control protocol; a signaling protocol for Internet telephony. SIP can establish sessions for features such as audio/videoconferencing, interactive gaming, and call forwarding to be deployed over IP networks, thus enabling service providers to integrate basic IP telephony services with Web, e-mail, and chat services. In addition to user authentication, redirect, and registration services, the SIP server supports traditional telephony features such as personal mobility, time-of-day routing, and call forwarding based on the geographical location of the person being called.

Unlicensed Mobile Access (UMA): The technology that provides access to global system for mobile communications (GSM) and general packet radio service (GPRS) mobile services over unlicensed spectrum technologies, including Bluetooth and 802.11. By deploying UMA technology, service providers can enable subscribers to roam and handover between cellular networks and public and private unlicensed wireless networks using dual-mode mobile handsets.

Voice Over Internet Protocol (VoIP): The routing of voice conversations over the Internet or any other IP-based network. The voice data flows over a general-purpose packet-switched network, instead of traditional dedicated, circuit-switched telephony transmission lines. Voice over IP traffic might be deployed on any IP network, including those lacking a connection to the rest of the Internet, for instance on a private building-wide LAN.

Contractual Obligations between Mobile Service Providers and Users

Robert Willis
Lakehead University, Canada

Alexander Serenko
Lakehead University, Canada

Ofir Turel
McMaster University, Canada

INTRODUCTION

The purpose of this chapter is to discuss the effect of contractual obligations between users and providers of mobile services on customer loyalty. One of the unique characteristics of mobile commerce that distinguishes it from most other goods and services is the employment of long-term contractual obligations that users have to accept to utilize the service. In terms of over-the-counter products, sold in one-time individual transactions in well-established markets, a strong body of knowledge exists that suggests that businesses may enhance loyalty through the improvement of quality and customer satisfaction levels. With respect to mobile commerce, however, this viewpoint may not necessarily hold true given the contractual nature of business-customer relationships.

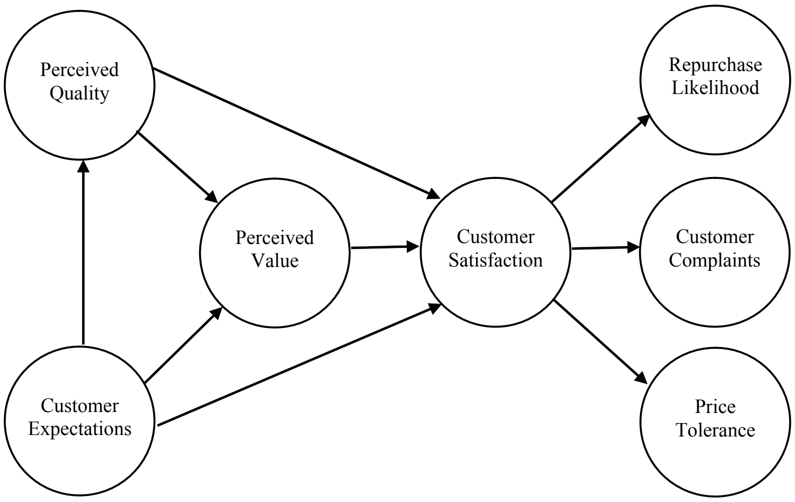
In the case of mobile computing, it is suggested that loyalty consists of two independent yet correlated constructs that are influenced by different factors: repurchase likelihood and price tolerance. Repurchase likelihood is defined as a customer's positive attitude towards a particular service provider that increases the likelihood of purchasing additional services or repurchasing the same services in the future (e.g., after the contract expires). For example, when people decide to purchase a new mobile phone, they are free to choose any provider they want. In other words, repurchase likelihood is not affected by contractual obligations. In contrast, price tolerance corresponds to a probability of staying with a current provider when it increases or a competitor decreases service charges. In this situation, individuals have to break the existing contractual obligations. Currently, there is empirical evidence to suggest that the discussion above holds true in terms of mobile computing. However, there are few well-documented works that explore this argument in depth. This article attempts to fill that void.

This article will present implications for both scholarship and practice. In terms of academia, it is believed that researchers conducting empirical investigations on customer loyalty with mobile services should be aware of the two independent dimensions of the business-customer relationship and utilize appropriate research instruments to ensure the unidimensionality of each construct. With regards to practice, it is suggested that managers and marketers be aware of the differences between repurchase likelihood and price tolerance, understand their antecedents, and predict the consequences of manipulating each one. It is noted that overall loyalty is not the only multidimensional constuct in mobile commerce. Recently, it was emperically demonstrated that perceived value of short messaging services is a second-order construct that consists of several independent yet correlated dimensions (Turel et al., 2007).

Theoretical separation of the overall loyalty construct into two dimensions has been already empirically demonstrated in three independent mobile commerce investigations. First, Turel and Serenko (2006) applied the American customer satisfaction model (ACSM) to study mobile services in North America. By utilizing the original instrument developed by Fornell, Johnson, Anderson, Cha, and Bryant (1996), they discovered a low reliability of the overall satisfaction construct, and found that the correlation between two items representing price tolerance and one item reflecting repurchase likelihood was only 0.21 ($p<0.01$, N=204). Second, Turel et al. (2006) adapted the ACSM to study the consequences of customer satisfaction with mobile services in four countries (Canada, Finland, Israel, and Singapore), and reported that the correlation between price tolerance and repurchase likelihood was 0.20 ($p<0.01$, N=736). Third, Yol, Serenko, and Turel (2006) analyzed the ACSM with respect to mobile services in the U.S. and again found the same correlation to be 0.45 ($p<0.01$, N=1,253). All these correlations fall into the small-to-medium range, and two of them are beyond the lowest cut-off value of 0.35 for item-to-total correlation (Nunnally & Bernstein, 1994). The statistical significance of these correlations is explained by large sample sizes. Therefore, it is impossible to design a single unidimensional construct in mobile commerce research consisting of both price tolerance

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. The American Customer Satisfaction Model (adapted from Fornell et al., 1996)



and repurchase likelihood. In all of these studies, most users had long-term contractual obligations with their respective mobile service provider that confirms the validity of the aforementioned conceptual discussion.

To better understand the customer loyalty concept in light of contractual obligations, this article briefly describes the American customer satisfaction model (ACSM), and then discusses the concepts of price tolerance and repurchase likelihood. Finally, it presents a summary which outlines implications for research and practice.

THE AMERICAN CUSTOMER SATISFACTION MODEL

The mobile telephony market continues to be one of the fastest growing service sector markets, creating a fiercely competitive industry environment (Kim & Yoon, 2005). As has happened in other, subscription-based mobile service industries, the nature of this competition has changed from the acquisition of new customers to the retention of existing customers and the luring away of competitors' customers. This last strategy is known in the industry as *outbound churn* or, more simply, as *churn*. Given the increasing penetration of mobile computing devices and the maturation of the market, avoiding churn and maximizing customer loyalty has become a primary concern for wireless providers. The first step in minimizing churn in a company's customer base is to understand its root causes.

The determinants of churn may be estimated by the adapted American Customer Satisfaction Model (see Figure 1). The original model suggests that satisfaction affects overall customer loyalty, where loyalty is a unidimensional construct that consists of price tolerance (i.e., the probability of staying with the current provider if it increases prices or if a competitor decreases prices) and repurchase likelihood (i.e., the probability of purchasing the same service again). At the same time, several recent works suggest that these loyalty dimensions are distinct yet correlated because of the contractual nature of the customer-service provider relationship.

Customer loyalty is one of the major constructs in marketing, and a large part of a marketing manager's effort is aimed at creating and maintaining loyalty among an organization's customer base. The significance of loyalty comes from the positive impact it has on the operations of the company in terms of customer retention, repurchase, long-term customer relationships, and company profits (Caruana, 2004). In other words, loyalty is a primary factor in reducing churn.

The notion of switching costs affecting loyalty has been recognized and researched by several professional and academic disciplines, including marketing, economics, and strategy. "Switching costs are generally defined as costs that deter customers from switching to a competitor's product or service" (Caruana, 2004, p. 256). For managers and researchers, it is important to understand the concepts of switching costs and customer loyalty, and to clearly identify both their dimensions and their interaction.

PRICE TOLERANCE

Switching costs are generally defined as one-time costs facing the consumer/buyer of switching from one supplier to another (Porter, 1980; Burnham, Frels, & Mahajan, 2003). Several researchers have identified various attributes or types of switching costs (e.g., Thibault & Kelly, 1959; Klemperer, 1987; Guiltnan, 1989; Burnham et al., 2003; Hu & Hwang, 2006); however, for the purposes of this article, switching costs are broadly categorized as three types: transaction, learning, and contractual. Transaction costs are costs incurred when a consumer begins a relationship with a provider and includes the costs associated with ending that relationship or terminating an existing relationship. Learning costs are associated with the effort required by the consumer to achieve the same level of knowledge and comfort acquired using a particular supplier's product, but which may not be transferable to similar/same products of other suppliers. Additionally, the notion of learning costs incorporates the implicit switching costs associated with decision biases, risk aversion, and market knowledge/familiarity. Learning what one's options are, what the relative competitive position of all suppliers are, and other such knowledge involves learning costs that will be differentially valued by individuals. In the case of mobile services, the switching costs are created by a service provider that requires customers to sign a long-term contract. If a customer wants to switch to another provider, he or she will have to pay a penalty to the current provider. As such, contractual costs are those costs that are directly provider induced in order to penalize churn and which are intended to prevent poaching of customers by other suppliers. With respect to the American customer satisfaction model, switching costs directly affect price tolerance. The ACSM survey instrument presents two questions: (1) by how much their current provider should increase its current prices in order for them to switch to a competitor, and (2) by how much a competitor should reduce its prices in order for them to switch. Peoples' answers to these questionnaire items are greatly affected by the direct switching costs they incur, such as a penalty.

Consistent with this proposition, Weiss and Anderson (1992) found that switching costs are a major consideration when consumers are making a churn decision, and that these costs (barriers) tend to reduce customers' churn behavior. These findings were further supported by research done by Jones, Mothersbaugh, and Beatty (2000). Burnham et al. (2003) suggested that switching costs are negatively correlated with a customer's intention to churn: the higher the costs, the lower the intention to switch. As Hu and Hwang (2006) point out, "the industry remains in a state of dynamic competition" (p. 75) and providers continue implementing flexible offerings that are aimed at reducing consumers' churn behavior. Shapiro and Varian (1999) found that *perceived* switching costs—which incorporate all of the explicit costs as well as the implicit costs discussed above—act as barriers to churn behavior. They suggest consumers will weigh the benefits of switching against the actual and psychological costs when considering churning.

Overall, the discussion above demonstrates that the concept of switching barriers has its own unique dimensions. In terms of the American customer satisfaction model applied in the context of mobile services, it is believed that two items pertaining to the customer switching behavior (conceptualized as price tolerance in the model) reflect a unique latent variable entitled *price tolerance.*

REPURCHASE LIKELIHOOD

The notion of overall customer loyalty has changed in both breadth and depth over the years in which it has been studied by academics and practitioners alike. The breadth of its definition is demonstrated by the multiplicity of areas that are examined, such as brand, product, vendor, or service loyalty. Initial research was primarily focused on brand loyalty, and mostly examined the behavioral aspects of the construct. In this view, Newman and Werbel (1973) defined customer loyalty as the repurchase of a brand that only considered that brand and which involved no brand-related information seeking.

Day (1969) was one of the first researchers to highlight the role of a positive attitude in the construct of loyalty. Following this line of reasoning, which incorporated both the behavioral and attitudinal conceptions of loyalty, operationalization of the construct of customer loyalty involved combining the aspects of purchasing a particular brand together with an affective attitudinal measure, whether that measure used a single scale or multi-scale items. With regards to the American customer satisfaction model, the discussion above relates to the unique dimension of loyalty as *repurchase likelihood,* or the probability of buying new services from the current provider when these purchases are not affected by prior contractual obligations, for example, when a contract has expired.

PRICE TOLERANCE AND REPURCHASE LIKELIHOOD

The literature—and intuition—suggests that higher switching costs are positively related to price tolerance—that is, that higher switching costs compel customers to remain loyal. Fornell et al. (1996) were among the first to include switching costs by adding them to the construct of customer satisfaction in the reflection of customer loyalty. In the ACSM, all items (i.e., two pertaining to price tolerance and one relating to repurchase likelihood) were believed to reflect overall loyalty. A number of subsequent studies demonstrated the unidimensionality of this construct. However, in the context of mobile services when high switching costs exist, unidimensionality does not apply. As such, it is suggested that, based on the theoretical rationale as well as empirical studies cited earlier, loyalty should be analyzed along two distinct dimensions: price tolerance and repurchase likelihood.

In terms of prior empirical research, Jones and Sasser (1995) included switching costs as one factor or competitiveness: since high switching costs discourage churning, they reduce the incentive for firms to compete. Bateson and Hoffman (1999) similarly suggest that customer satisfaction and switching costs are the primary influencers of loyalty. More recent studies have shown that switching costs have a direct and strong influence on the re-purchase decision (customer loyalty) in all markets, for example France (Lee, Lee, & Feick, 2001), Korea (Kim & Yoon, 2005), Australia (Caruana, 2004), Taiwan (Hu & Hwang, 2006), and Turkey (Aydin, Özer, & Arasil, 2005).

Jones et al.'s (2000) study examined the role of switching costs (barriers) in customer retention for services. They found that although core-service satisfaction was a primary issue in retention, switching factors in the form of interpersonal relationships, direct and indirect costs, and the perceived benefits of potential alternatives were also important. As such, these factors represented different unique dimensions of the overall loyalty concept. This supports the notion, outlined above, that loyalty of mobile service users must be considered as multidimensional and not simply as direct, contractual costs.

IMPACTS FOR MANAGERS AND RESEARCHERS

The findings of the many studies in the area show support for the intuitive link between higher switching costs and greater levels of customer loyalty (or at least, retention). More importantly, they also provide a greater understanding of the interaction between switching costs and loyalty, and refine the model that has, to date, served as a guide to management of mobile phone companies.

Management of mobile phone companies must understand the complexity and multidimensionality of the concepts of switching costs that directly influence price tolerance and repurchase likelihood that is not affected by contractual obligations. They must also understand that switching costs affect customer loyalty not solely through the contractual cost component of switching costs, but also through the learning and transaction cost components. A customer's, or potential customer's, belief that he or she will end up with a 'bad deal' financially in switching to a new provider—and that assessment will include all of the implicit as well as explicit costs—is the most important issue in the churn decision. This highlights the point that managing customer relationships, so that they remain positive, acts to keep the customer attached, whether this is a result of satisfaction outweighing perceived benefits or simply of customer inertia (Burnham et al, 2003; Caruana, 2004). It also highlights the need for poaching strategies to emphasize not only the financial benefits, but the relational benefits as well (Hu & Hwang, 2006). It should be noted that existing studies point out that one of the primary issues affecting the learning cost component has been the lack of time to undertake a complete comparison of the many offerings in the market. Additionally, providers have tended in the past to couch their offerings in terms that vary widely from their competitors', thus introducing a level of uncertainty and confusion in the minds of the analyzing consumer (Hu & Hwang, 2006). These factors are becoming less and less viable as consumers turn to the Internet for their purchasing information and guidance, and as consumers demand—and get—a certain level of standardization in the offerings of providers in the market, whether that standardization comes from the providers themselves or from organizations that perform such analyses and offer them to the consuming (Internet- or magazine-based) consumer. Additionally, the increasing homogeneity of pricing strategies and service packages will lead to a lessening of the impact of explicit (transaction and contractual) switching costs on the churn decision (Hu & Hwang, 2006). Thus, management needs to concentrate on customer relationships. Swartz (2000) quotes two senior executives in the industry:

> *If service is poor, then customers will pay any cancellation fees to get rid of the service and choose another provider….*
>
> *You have to look at your reasons for churn…You can't use a contract to make up for poor service. If your service is poor, you can lock them in for a year…but they're gone the minute month 13 rolls around.*

More research needs to be done on the notion of overall loyalty as a multidimensional construct. Are there positive barriers, such as interpersonal relationships, as well as negative? What relative influence on customer satisfaction do core and non-core services have? How sensitive are costs as barriers? Research into whether or not there are services that are perceived as having low barriers as opposed to services that are perceived as having high barriers within the market offerings would help refine our understanding of the role of various costs.

REFERENCES

Aydin, S., Özer, G., & Arasil, Ö. (2005). Customer loyalty and the effect of switching costs as a modifier variable: A case in the Turkish mobile phone market. *Marketing Intelligence and Planning, 23*(1), 89-103.

Bateson, J. E. G., & Hoffman, K. D. (1999). *Managing services marketing, text and readings* (4th ed.). Fort Worth, TX: Dryden Press.

Burnham, T. A., Frels, J. K., & Mahajan, V. (2003). Consumer switching costs: A typology, antecedents and consequences. *Journal of the Academy of Marketing Science, 31*(2), 109-126.

Caruana, A. (2004). The impact of switching costs on customer loyalty: A study among corporate customers of mobile telephony. *Journal of Targeting, Measurement and Analysis for Marketing, 12*(3), 256-268.

Day, G. S. (1969). A two dimensional concept of brand loyalty. *Journal of Advertising Research, 9*(3), 29-36.

Fornell, C. (1992). A national consumer satisfaction barometer: The Swedish experience. *Journal of Marketing, 56*(1), 6-21.

Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American Customer Satisfaction Index: Nature, purpose, and findings. *Journal of Marketing, 60*(4), 7-18.

Guiltnan, J. P. (1989). A classification of switching costs with implications for relationship marketing. In T. L. Childers & R. P. Bagozzi (Eds.), *Proceedings of the Winter Educators' Conference: Marketing Theory and Practice* (pp. 216-220), Chicago.

Hu, A.W.-L., & Hwang, I.-S. (2006). Measuring the effects of consumer switching costs on switching intention in Taiwan mobile telecommunications services. *Journal of American Academy of Business, 9*(1), 75-85.

Jones, M. A., Mothersbaugh, D. L., & Beatty, S. E. (2000). Switching barriers and repurchase intentions in services. *Journal of Retailing, 76*(2), 259-274.

Jones, T. O, & Sasser, W. E. (1995). Why satisfied customers defect. *Harvard Business Review, 73*(1), 88-99.

Kim, H.-S., & Yoon, C.-H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28,* 751-756.

Klemperer, P. (1987). Markets with consumer switching costs. *The Quarterly Journal of Economics, 102,* 375-394.

Lee, J., Lee, J., & Feick, L. (2001). The impact of switching costs on customer satisfaction-loyalty link: Mobile phone service in France. *Journal of Services Marketing, 15*(1), 35-48.

Newman, J. W., & Werbel, R. A. (1973). Multivariate analysis of brand loyalty for major household appliances. *Journal of Marketing Research, 10*(4), 404-409.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oliver, R. L. (1996). *Satisfaction: A behavioral perspective on the consumer.* New York: McGraw-Hill.

Porter, M. E. (2003). *Competitive strategy: Techniques for analyzing industries and competitors.* New York: MacMillan.

Serenko, A., Turel, O., & Yol, S. (2006). Moderating roles of user demographics in the American customer satisfaction model within the context of mobile services. *Journal of Information Technology Management, 17*(4): in-press.

Swartz, N. (2000). Reconsidering contracts. *Wireless Review, 17*(4), 48-52.

Thibault, J. W., & Kelley, H. H. (1959). *The social psychology of groups.* New York: John Wiley & Sons.

Turel, O., & Serenko, A. (2006). Satisfaction with mobile services in Canada: An empirical investigation. *Telecommunications Policy, 30*(5-6), 314-331.

Turel, O., Serenko, A., & Bontis, N. (2007). User acceptance of wireless short messaging services: Deconstructing perceived value. *Information & Management, 44*(1), 63-73.

Turel, O., Serenko, A., Detlor, B., Collan, M., Nam, I., & Puhakainen, J. (2006). Investigating the determinants of satisfaction and usage of mobile IT services in four countries. *Journal of Global Information Technology Management, 9*(4), 6-27.

Weiss, A.M., & Anderson, E. (1992). Converting from independent to employee sales forces: The role of perceived switching costs. *Journal of Marketing Research, 29*(1), 101-115.

KEY TERMS

American Customer Satisfaction Model: The original model suggests that satisfaction affects the overall customer loyalty, where loyalty is a unidimensional construct that consists of price tolerance (i.e., the probability of staying with the current provider if it increases prices or if a competitor decreases prices) and repurchase likelihood (i.e., the probability of purchasing the same service again). If the customer's expectations of product quality, service quality, and price are exceeded, a firm will achieve high levels of customer satisfaction and will create *customer delight*. If the customer's expectations are not met, customer dissatisfaction will result. And the lower the satisfaction level, the more likely the customer is to stop buying from the firm.

Churn: This refers to the notion that a company will, over any given period of time, lose existing customers and gain new customers. Churn is, currently, mostly created by the luring away of competitors' customers.

Customer Loyalty: The notion that a customer will continue to use a particular brand or product; the behavior customers exhibit when they make frequent repeat purchases of a brand or product.

Price Tolerance: The extent to which price is an important criterion in the customer's decision-making process; thus a price-sensitive customer is likely to notice a price rise and switch to a cheaper brand or supplier.

Repurchase Likelihood: The probability of buying new services from the current provider when these purchases are not affected by prior contractual obligations, for example, when a contract has expired.

Switching Cost: One-time cost facing the consumer/buyer of switching from one supplier to another. Switching costs are composed of transaction costs (costs incurred when a consumer begins a relationship with a provider, and includes the costs associated with ending that relationship or terminating an existing relationship), learning costs (costs associated with the effort required by the consumer to achieve the same level of knowledge and comfort acquired using a particular supplier's product, but which may not be transferable to the same/similar products of other suppliers), and contractual costs (costs that are directly provider induced in order to penalize churn and which are intended to prevent poaching of customers by other suppliers).

Convergence Technology for Enabling Technologies

G. Sivaradje
Pondicherry Engineering College, India

I. Saravanan
Pondicherry Engineering College, India

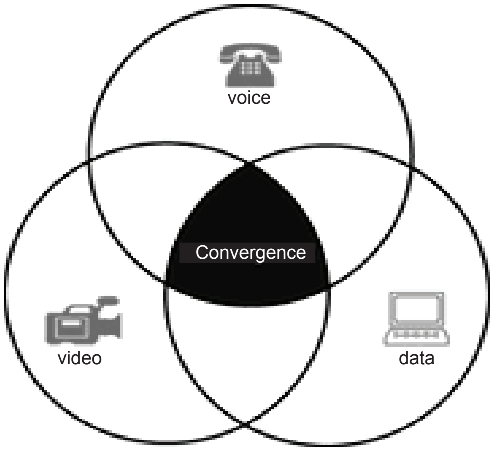
P. Dananjayan
Pondicherry Engineering College, India

INTRODUCTION

Today, we find a large number of wireless networks based on different radio access technologies (RATs). Every existing RAT has its own merits. Now the focus is turned towards the next-generation communication networks (Akyildiz, Mohanty, & Xie, 2005), which will seamlessly integrate various existing wireless communication networks, such as wireless local area networks (WLANs, e.g., IEEE 802.11 a/b/g and HIPERLAN/2), wireless wide area networks (WWANs, e.g., 1G, 2G, 3G, IEEE 802.20), wireless personal area networks (WPANs, e.g., Bluetooth, IEEE 802.15.1/3/4), and wireless metropolitan area networks (WMANs, e.g., IEEE 802.16) to form a converged heterogeneous architecture (Cavalcanti, Agrawal, Cordeiro, Xie, & Kumar, 2005). Seamless integration does not mean that the RATs are converged into a single network. Instead the services offered by the existing RATs are integrated as shown in Figure 1.

Convergence technology is a technology that combines different existing access technologies such as cellular, cordless, WLAN-type systems, short-range wireless connectivity, and wired systems on a common platform to complement each other in an optimum way and to provide a multiplicity of possibilities for current and future services and applications to users in a single terminal. After creating a converged heterogeneous architecture, the next step is to perform a common radio resource management (RRM) (Magnusson, Lundsjo, Sachs, & Wallentin, 2004). RRM helps to maximize the use of available spectrum resources, support mixed traffic types with different QoS requirements, increase trunking capacity and grade of service (GoS), improve spectrum usage by selecting the best RAT based on radio conditions (e.g., path loss), minimize inter-system handover latency, preserve QoS across multiple RATs, and reduce signaling delay. A typical converged heterogeneous architecture (Song, Jiang, Zhuang, & Shen, 2005) is shown in Figure 2.

Figure 1. Convergence of services



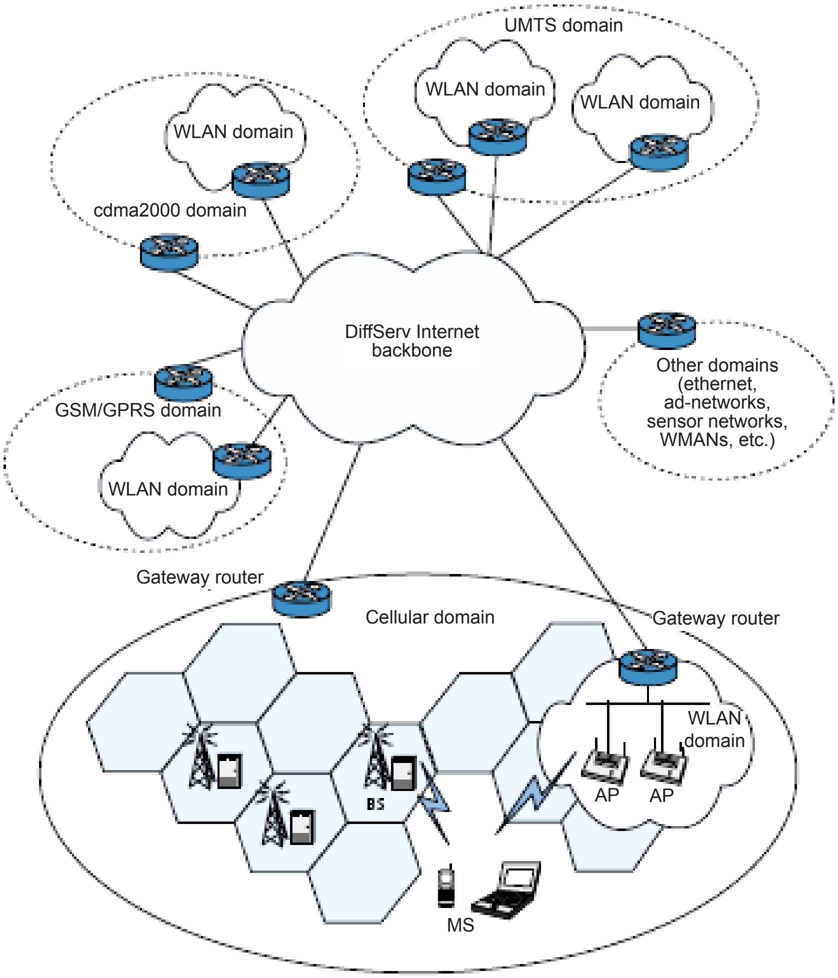
CHALLENGES

The integration of different networks to provide services as a single interworking network requires many difficult challenges to be addressed. Because existing networks do not have fair RRM, the major challenge that needs to be addressed has to be mobility management. The heterogeneous network architecture will be based on IP protocol that will enhance the interoperability and flexibility. IETF Mobile IP protocol is used to support macro mobility management. But both IP protocol and mobile IP protocol (Pack & Choi, 2004; Montavont & Noel, 2002) was not basically designed to support the real-time applications. So, during the handoff between systems, users will experience the service discontinuity, such as long service time gap or network disconnection. Besides this service discontinuity, the different service characteristics of these interworked networks may degrade the quality of service (QoS).

Some of the other challenges include topology and routing, vertical handoff management, load balancing, unified

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Converged heterogeneous network architecture


UMTS domain
WLAN domain
WLAN domain
WLAN domain
cdma2000 domain
DiffServ Internet backbone
Other domains (ethernet, ad-networks, sensor networks, WMANs, etc.)
GSM/GPRS domain
WLAN domain
Gateway router
Cellular domain
Gateway router
WLAN domain
AP
AP
BS
MS

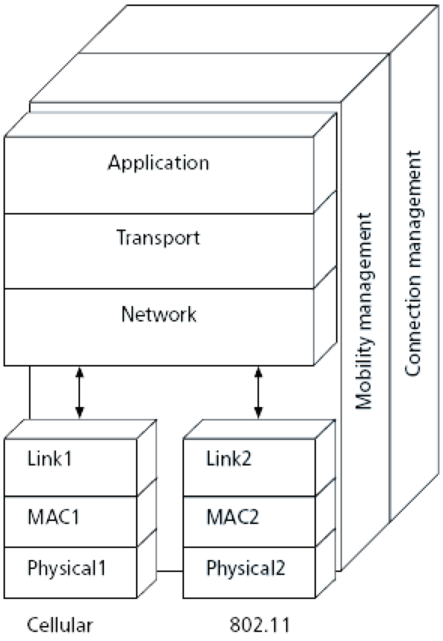
accounting and billing, and last but not least the protocol stack of mobile station (MS), which should contain various wireless air-interfaces integrated into one wireless open terminal so that same end equipment can flexibly work in the wireless access domain as well as in the mobile cellular networks.

PROTOCOL STACK

In a homogeneous network, all network entities run the same protocol stack, where each layer has a particular goal and provides services to the upper layers. The integration of different technologies with different capabilities and functionalities is an extremely complex task and involves issues at all the layers of the protocol stack. So in a heterogeneous environment, different mobile devices can execute different protocols for a given layer. For example, the protocol stack of a dual-mode MS is given in Figure 3.

This protocol stack consists of multiple physical, data link, and medium access control (MAC) layers, and network, transport, and application layers. Therefore, it is critical to select the most appropriate combination of lower layers (link, MAC, and physical) that could provide the best service to the upper layers. Furthermore, some control planes such as mobility management and connection management can be added. These control planes can eventually use information from several layers to implement their functionalities. The network layer has a fundamental role in this process, since it is the interface between available communications interfaces (or access technologies) that operate in a point-to-point fashion, and the end-to-end (transport and application) layers. In other words, the task of the network layer is to provide a uniform substrate over which transport (e.g., TCP and UDP), and application protocols can efficiently run, independent of the access technologies used in each of the point-to-point links in an end-to-end connection. Although there are issues in all layers, the network layer has received more attention than

Figure 3. Protocol stack of a dual-mode MS



any other layer, and little integration-related work has been done at the lower layers. Indeed, integrated architectures are expected not to require modifications at the lower layers so that different wireless technologies can operate independently. However, this integration task is extremely complex, and it requires the support of integration architecture in terms of mobility and connection management. Seamless handoffs for "out of coverage" terminals and resource management can be provided by the two control planes.

ROUTING ISSUES

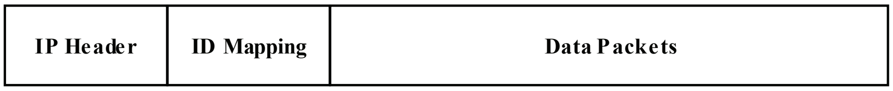
All RATS in the integrated architecture is considered as IPv6-based networks, and each element in the internetworking networks has a distinct ID number corresponding to the network routing address (Liu & Zhou, 2004). The infrastructure of a network is mapped into IPv6 addresses as shown in Figure 4. For example, the mapping of infrastructure of cellular network and IEEE 802.11 WLAN are shown in Figures 5 and 6. WLAN is given some reservation IDs, so that they can be utilized by mobile nodes under MANET mode.

VERTICAL HANDOFF MANAGEMENT

Vertical handoff is the handoff between different RATs. The major challenge in vertical handoff is that it is difficult to support a seamless service during inter-access network handoff (Wu, Banerjee, Basu, & Das, 2005; Ma, Yu, Leung, & Randhawa, 2004). The service interworking architecture and procedures, the way to provide the network and user securities, the control scheme for minimizing performance decrease caused by different service data rates, and the interworking network detection and selection methods are typical problems and to be addressed to provide stable and continuous services to users.

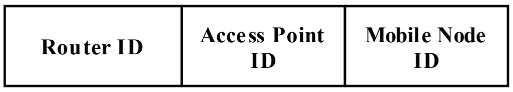
Unlike in the homogeneous wired networks, providing QoS for integrated architecture has some fundamental bottlenecks. This is because each radio access technology has different transmission-rate capacity over the radio interfaces, therefore the handoff between the two systems makes the maintenance of QoS connection very hard. For example, WLAN can provide a transmission speed from 11Mb/s up to 54Mb/s theoretically, while UMTS has only 144kb/s at vehicular speed, 384 kb/s at pedestrian speed, and 2 Mb/s when used indoors. If we keep the QoS resource assigned by UMTS to a connection that is actually in a WAN hotspot, the advantage of the high speed of the WLAN is not fully taken. On the other hand, if we use a WLAN parameter for a station in the UMTS network, the connection may not be admitted at all (Zhang et al., 2003). Therefore, to maintain a sensible QoS framework, one has to consider the significant difference transmission capacity between two systems especially when user handover takes place.

Figure 4. Mapping infrastructure into IPv6 format



IP Header	ID Mapping	Data Packets

Figure 5. Mapping infrastructure of IEEE 802.11 WLAN into IPv6 address



Router ID	Access Point ID	Mobile Node ID

Figure 6. Mapping infrastructure of cellular network into IPv6 address

Network ID	RNC ID	Base Station ID	Mobile Node ID

APPLICATIONS

Convergence technology gives the possibility to combine audio and video, data, graphics, slides and documents, and Internet services in any way you like, so as to maximize the effectiveness of the communication. Integrating all traffic types enables more versatile and efficient ways of working, not just internally to the organization, but externally to customers, partners, and suppliers. It also creates a multi-system environment where a single service could be offered at different speeds at different locations/times via separate systems. The flexibility of convergence technology provides many applications and services to the user community. Some of the applications are:

- **Find-Me-Follow-Me:** This is a customizable service that makes it easy for callers to 'find' a user. Using a Web portal customers can choose how incoming calls should be handled. Options include ringing multiple phones simultaneously, or picking the order of phones to ring sequentially. Ubiquity's SIP A/S is used to dial out, in parallel or sequentially, to the user's contact numbers. Using IVR, the user can then accept the call or forward it to voicemail.
- **InfoChannels:** This is a multimedia content subscription application that pushes information and entertainment to users in real time. Users subscribe to content services through a Web portal, and new content is delivered to their designated device (mobile phone, PDA, PC browser) as soon as it is available.
- **Rich Media Conferencing:** Speak conference director is a highly scalable, carrier-class, IP conferencing application that enables conferencing service providers (CSPs) to offer hosted audio and Web conferencing services. This easy-to-use, browser-based solution offers a complete conferencing application feature set, as well as a Web portal for scheduling, initiating, managing, and terminating multi-party conferences.

Some of the services that the convergence provides to the user community are:

- **Unified Messaging:** Same inbox handling data, voice and fax.
- **Hosted IP Voice:** A complete, outsourced telephone service offering all PBX-type features.
- **IP Fax:** Delivery of e-mail to fax and fax to e-mail in a large number of countries.
- **IP Telephony:** A combination of quality transmission globally across the WAN and the LAN, with tailored consulting and end-to-end support.
- **Voice for IP VPN:** Integrated voice and data transmission, using a specific voice.
- **Video for IP VPN:** Point-to-point video transmission over the IP VPN network, using a specific class of service, called RT Vi.
- **Virtual Contact Center Services:** Optimization of agent resources while reducing costs, by allowing the routing of calls based on the agent's skills.
- **Voiceover Wi-Fi:** Full corporate mobility with a converged voice and data wireless solution.

CONCLUSION

This article provides features about convergence technology. The convergence of all existing networks will provide access to all available services using a single-user terminal. But there are many challenges to be addressed in converging the networks. In spite of converging the networks, management of the converged network is more challengeable. This article illustrates some of the challenges, and many are still open issues. Considering all the factors discussed, convergence technology is going to provide future flexibility to the wireless communication world. The complexity of this interesting technology must be addressed in the near future.

REFERENCES

Akyildiz, I. F., Mohanty, S., & Xie, J. (2005). A ubiquitous mobile communication architecture for next-generation heterogeneous wireless systems. *IEEE Radio Communications, 43*(6), S29-S36.

Cavalcanti, D., Agrawal, D., Cordeiro, C., Xie, B., & Kumar, A. (2005). Issues in integrating cellular networks, WLANS, and MANETs: A futuristic heterogeneous wireless network. *IEEE Wireless Communications, 12*(3), 30-41.

Liu, C., & Zhou, C. (2004). HCRAS: A novel hybrid internetworking architecture between WLAN and UMTS cellular networks. In *Proceedings of IEEE 2004* (pp. 374-379).

Ma, L., Yu, F., & Leung, V. C. M., & Randhawa, T. (2004). A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wireless Communication, 11*(4), 44-51.

Magnusson, P., Lundsjo, J., Sachs, J., & Wallentin, P. (2004). Radio resource management distribution in a Beyond 3G Multi-Radio Access architecture. In *Proceedings of the IEEE Communications Society, Globecom* (pp. 3372-3477).

Montavont, N., & Noel, T. (2002). Handover management for mobile nodes in IPv6 networks. *IEEE Communications Magazine, 40*(8), 38-43.

Pack, S., & Choi, Y. (2004). A study on performance of hierarchical mobile IPv6 in IP-based cellular networks. *IEICE Transactions on Communication, E87-B*(3), 546-551.

Song, W., Jiang, H., Zhuang, W., & Shen, X. (2005). Resource management for QoS support in cellular/WLAN interworking. *IEEE Network, 19*(5), 12-18.

Wu, W., Banerjee, N., Basu, K., & Das, S. K. (2005). SIP-based vertical handoff between WWANS and WLANS. *IEEE Wireless Communications, 12*(3), 66-72.

Zhang, Q., Guo, C., Guo, Z., & Zhu, W. (2003). Efficient mobility management for vertical handoff between WWAN and WLAN. *IEEE Communication Magazine, 41*(11), 102-108.

KEY TERMS

Communication Network: Network of telecommunications links arranged so that messages may be passed from one part of the network to another over multiple links.

Grade of Service (GoS): A measurement of the quality of communications service in terms of the availability of circuits when calls are to be made. Grade of service is based on the busiest hour of the day and is measured as either the percentage of calls blocked in dial access situations or average delay in manual situations.

Heterogeneous Network: A network that consists of workstations, servers, network interface cards, operating systems, and applications from many vendors, all working together as a single unit.

Radio Access Technology (RAT): Technology or system used for the cellular system (e.g., GSM, UMTS, etc.).

Wireless Local Area Network (WLAN): Wireless network that uses radio frequency technology to transmit network messages through the air for relatively short distances, like across an office building or college campus.

Wireless Metropolitan Area Network (WMAN): A regional wireless computer or communication network spanning the area covered by an average to large city.

Wireless Personal Area Network (WPAN): Personal, short-distance area wireless network for interconnecting devices centered around an individual person's workspace.

Wireless Wide Area Network (WWAN): Wireless network that enables users to establish wireless connections over remote private or public networks using radio, satellite, and mobile phone technologies instead of traditional cable networking solutions like telephone systems or cable modems over large geographical areas.

Cooperative Caching in a Mobile Environment

Say Ying Lim
Monash University, Australia

INTRODUCTION

There has been a rapid recent development in the usage of mobile devices, such as personal digital assistant (PDA), mobile notebooks and other mobile electronic devices. This has been increasingly in demand due to new smaller designs and easy to carry around features. The mobile computing technology has pushed the uses of mobile device even more by providing the ability for mobile users to access information anytime, anywhere. However, mobile computing environments have certain limitations, such as short battery power and limited storage and involving communication cost and bandwidth limitations (Kossman et al., 2001). Hence, it is of great interest to provide efficient query processing with quick response rate and with low transfer cost. Due to the inherent factors like low bandwidth and low reliability of wireless channels in the mobile computing environment, it is therefore important for a mobile client to cache its frequently accessed database items into its local storage (Chan, Si, & Leong, 2001).

Caching is a key strategy in improving data retrieval performance of mobile clients (Barbara & Imielinkski, 1994; Chow, Leong, & Chan, 2005). In order to retain the frequently accessed database items in the client's local memory, a caching mechanism is needed. By having the caching mechanisms, it allows a client to serve database queries at least partially during connection, which is an inherent constraint of the mobile environment. Thus, by having an effective caching mechanism in keeping the frequently accessed items, the more queries could be served in case of disconnection (Chan, Si, & Leong, 2001). The aim of caching is beneficial to the mobile environment, because having the data cached into the local memory can help future queries to be answered more quickly or to access the data faster, with low latency time and reduced start up delays that may be caused on the client side (Kara & Edwards, 2003). In addition to improving the access latency, it also helps to save power due to the ability to allow lower data transmission, as well as improvement in terms of data availability in situations of disconnection (Wu, Yu, & Chen, 1996; Lee, Xu, & Zheng, 2002).

In this article, we concentrate particularly on cooperative caching, which is basically a type of caching strategy that not only allows mobile clients to retrieve database items from the servers, but also from the cache in their peers.

BACKGROUND

The effect of having the ability to cache data is of great importance, particularly in the mobile computing environment. This is due to the reason that contacting the remote servers for data is expensive in the wireless environment and the vulnerability to frequent disconnection can further increase the communication costs (Leong & Si, 1997). Also, caching the frequently accessed data in the mobile devices will help reduce tune-in time and power consumption because requested data can be fetched for the cache without tuning into the communication channel for retrieval (Lee & Lee, 1999). Caching has also been proven as a significant technique in improving the performance of data retrieval in peer to peer network in helping to save bandwidth for each data retrieval that are made by the mobile clients (Joseph et al., 2005). The cached data are meant to support disconnected or intermitted connected operations. There are many different types of caching strategies that serve the purpose to improve query response time and to reduce contention on narrow bandwidth (Zheng, Lee, & Lee, 2004).

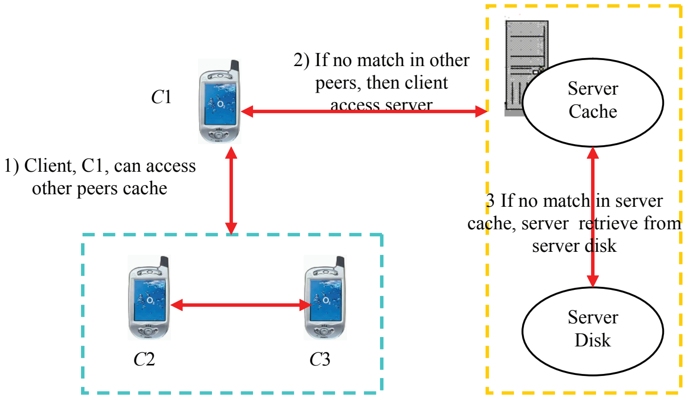
Regardless of which caching strategy one chooses, they have their own advantages and disadvantages in some way. In principle, caching can improve system performance in two ways: (a) It can eliminate multiple requests for the same data and (b) improve performance by off-loading the work. The first way can be achieved by allowing mobile clients to share data among each other by allowing them to access each other's cache within a reasonably boundary between them. The second way, however, can be demonstrated as in an example of a mobile user who is interested in keeping stock prices and caches them into his mobile device. By having copies in his own mobile device, he can perform his own data analysis based on the cached data without communicating directly to the server over the wireless channel.

COOPERATIVE CACHING

The cooperative caching is a kind of information sharing that was developed by the heavy influence and emergence of the robust yet reliable peer-to-peer (P2P) technologies, which allows mobile clients (Kortuem et al., 2001). This type of information sharing among clients in a mobile

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. An overview of cooperative caching architecture


C1
2) If no match in other peers, then client access server
Server Cache
1) Client, C1, can access other peers cache
3 If no match in server cache, server retrieve from server disk
C2
C3
Server Disk

environment has generally allowed the clients to directly communicate among themselves by being able to share cached information through accessing data items from the cache in their neighboring peers rather than having to rely on their communication to the server for each query request (Chow, Leong, & Chan, 2004).

There are several distinctive and significant benefits that cooperative caching brings to a mobile computing environment. These include improving access latency, reducing servers' workloads and alleviating point–to-point channel congestion. Though the benefits outweigh the drawbacks, there is still a main concern that cooperative caching may produce. This refers to the possibility of the increase in the communication overhead among mobile clients (Ku, Zimmermann, & Wan, 2005)

Framework Design of Cooperative Caching in a Mobile Environment

Several clients and servers are connected together within a wireless channel in a mobile environment. Basically, the clients which denote the mobile hosts are connected to each other wirelessly within a certain boundary among each other, and they can exchange information by allowing each other to access the other peers' cache (Cao, Yin, & Das, 2004).

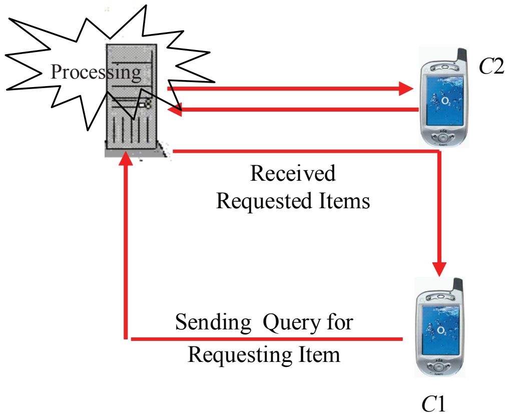
Figure 1 illustrates an example of the framework architecture design of the cooperative caching, which provides the ability for mobile clients to communicate among each other. If the client encounters a *local cache miss*, it will send a query to request, from its neighboring peers, to obtain a communication and the desired data from its peer's cache. Otherwise, it will be known as a *local cache hit* if the desired data exists in its local cache. As for trying to obtain data from its peers, if the desired data is available from its neighboring peers or if the other peers can turn in the requested data before it is broadcast on the channel, then it is known as a *global cache hit*; otherwise it is called a *global cache miss* and the client would have to wait for the desired data to arrive in the broadcast channel or access the server cache and, if that fails, then the server would retrieve the desired data from the disk (Sarkar & Hartman, 2001; Hara, 2002; Chow, Leong, & Chan, 2005).

As a summary, a mobile client can choose to either (a) retrieve data from the server directly by having a direct communication through issuing a query (Hu & Johnson, 2000) or (b) capture the data from a scalable broadcast channel (Su, Tassiulas, & Tsotras, 1999). These are known as a pull-base and push-base mechanism respectively. Further investigation on pull and push-based environments are made in the subsequent subsections.

Using Cooperative Caching in an On-Demand (Pull-Based) Environment

A pull-based environment refers to relating the use of traditional point–to-point scenario similar to client-server communication directly. It can also be known as an on demand query or server strategy whereby processing can be done on the server upon request sent by the mobile clients. Figure 2 illustrates an example of a pull-based architecture. It can be seen that the mobile client issues a direct query to the server over a dedicated channel to be processed. Processing takes

Figure 2. Example of an on demand environment system (pull-based system)



place in the server and, once the requested data items have been obtained, it will returned back to the client directly.

The main advantage of this system is that a client can issue query directly to the server and wait for the server to process and return the results according to the query being issued. Also, it is appropriate in situations where privacy is a major concern. On the other hand, the limitation is that this is not desirable in a mobile environment where there are limited resources to satisfy each individual client directly. Thus, this shows limitation when it comes to large-scale systems.

Example 1: Looking at Figure 3, suppose a shopper in a shopping complex wants to know which shops to visit by obtaining information from the store directories. Imagine that this client is denoted as *C*2 in Figure 3. If this shopper, *C*2, finds the target shop in its local cache, as she has previously visited this shopping complex and has cached the store directories in her mobile device. If this is the case, then she can just merely obtain the information from the cache. If not, then it will send request to its neighboring peers, which in this example are *C*1 and *C*3 since the boundary of the wireless transmission for *C*2 covers clients *C*1 and *C*3. So *C*2 can obtain the desired data from either *C*1 or *C*3. Assuming *C*3 has the data that *C*2 wanted, this means *C*2 can obtain the data from the cache in his peer, in this example *C*3. Otherwise, *C*2 would have to obtain it from the server directly.

Figure 3. Using cooperative caching in an on demand environment

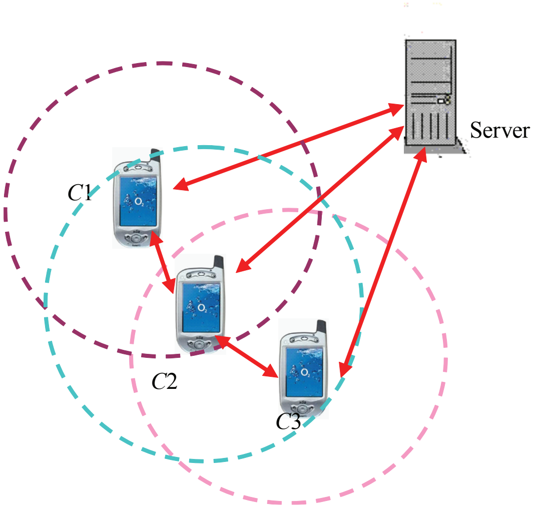
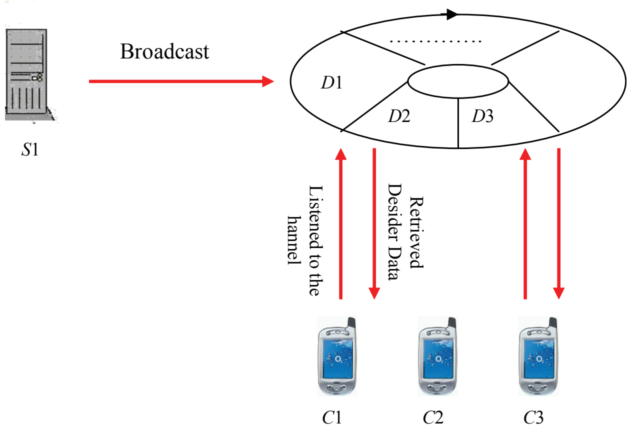


Figure 4. Example of a broadcast environment system (push-based system)



Using Cooperative Caching in a Broadcast (Push-Based) Environment

In this section, we focus on using cooperating caching in a broadcast environment, which can sometimes be referred to as a push-based system or also known as on-air strategy. In a broadcast environment, a mobile client is able to tune into the broadcast channel to retrieve the data that they want by having the server broadcast the data items into the air for the user to tune into (Waluyo, Srinivasan, & Taniar, 2005). Thus, each client can access a piece of data by waiting for its data to arrive. Figure 4 depicts an example of a broadcast environment. The main advantages of this broadcast environment in terms of its delivery mechanism is its higher throughput for data access for a larger number of clients because, with the absence of the communication contention between the clients requesting data, they are able to share the bandwidth more efficiently (Hara, 2002). Information broadcast appears to be an essential method for information dissemination for mobiles users, since it is able to broadcast to an arbitrary number of mobile users (Lee & Lee, 1999).

There are advantages and disadvantages of the broadcast environment. The strength would be its ability to disseminate data to an immense number of mobile clients. However, the greatest disadvantage lies in the way in which the data item are broadcast in a sequential way. This will lead to longer access latency if there is a substantial increase in the number of data items being broadcast. Also, by broadcasting the data item on the air, this means the mobile clients have to consume considerable amounts of power to listen to the broadcast channel until the target data items reaches its turn and appears in the broadcast channel.

In this section, we would like to illustrate the use of cooperative caching in the broadcast environment. It is always a benefit to allow the client to have access to more data in the mobile environment by improving the data availability and data accessibility (Cao, Yin, & Das, 2004). By having the clients share bandwidth together in the wireless channels; it is believed that cooperative caching is able to further save bandwidth for each data retrieval (Joseph et al., 2005). An example of an application utilizing this push-based mechanism can be a situation where, in the airport, up-to-date schedules could be broadcast and passengers with their mobile devices are able to receive and store the information.

Generally, there are a series of steps involved in cooperative caching in a broadcast environment. Basically, when a client issues a request to a particular data item on the broadcast channel, the issued request client would first check whether the desired data has been previously cached. If it has, then the answer would be immediately returned and succeed at once. However, if it has not, then it will secondly check to see if the response time in accessing the desired item that is cached by the neighboring peers is shorter than to wait for the item to appear on the broadcast channel. If it is, then the client will obtain it from its peers; otherwise it may be easier just to obtain it from the broadcast channel if it appears to be faster. Otherwise, it received a reply from other neighboring peers that cached the desired data item and the request would be completed and succeeded when the transmission of the item is completed. However, in case of a situation where no neighboring peers obtained the desired data, then the client would have to wait for the next broadcast period for the desired data item to appear (Hara, 2002).

FUTURE TRENDS

There have been several researches done in the area of cooperative caching in a mobile environment. Problems and limitations incurred in the usage of cooperative caching in the mobile environment have generated a lot of attention from researches in finding a good cache strategy that is specifically designed for use only in the mobile computing environment.

In the future, it is desirable to design techniques that can further improve the efficiency of caching in reducing the overhead of maintaining and identifying the cached items. Also exploring the semantics of the data items cached by the mobile clients is beneficial. Practical evaluations from various factors, such as communication and computational overhead, can be essential to determine which strategy appears most appropriate for a real mobile environment. It is also valuable to look into cache replacement and cache invalidation and addressing them in an environment where updates features takes place.

It is also advantageous to look further into incorporation of an effective replication scheme that may increase data availability and accessibility, as well as improve the average query latency. It would be useful to always take into account reducing the number of server requests and power consumption, as well as shortening the access latency as more neighboring peers increase. As far as most situations explain in this article, the focus does not take into account location dependency queries. Thus, it can be a good idea to consider location dependent queries, as well as accessing multiple non-collaborative servers instead of just obtaining data from a single server.

CONCLUSION

Caching appears to be a key factor in helping to improve the performance of answering queries in mobile environment. A new state-of-the-art information sharing known as cooperative caching allows mobile clients an alternative way to obtain desired data items. Clients can now have the ability to access data items from the cache in their neighboring peers' devices with the implementation of cooperative caching.

In this article, we have described issues of caching in a mobile environment by focusing mainly on cooperative caching in the mobile environment. Our discussion assumed a broadcast environment. This article serves as a helpful foundation for those who wish to gain preliminary knowledge about the usefulness and benefits of caching in the mobile environment, particularly cooperative caching.

REFERENCES

Barbara, D., & Imielinski, T. (1994, November). Sleepers and workaholics: Caching strategies in mobile environments. *MOBIDATA: An Interactive Journal of Mobile Computing*, *1*(1).

Chan, B. Y., Si, A., & Leong H. V. (1998). Cache management for mobile databases: Design and evaluation. In *Proceedings of the International Conference on Data Engineering (ICDE)* (pp.54-63).

Chow, C. Y., Leong, H. V., & Chan, A. T. S. (2005). Distributed group-based cooperative caching in a mobile environment. In *Proceedings of the 6th international conference on Mobile Data Management (MDM)* (pp. 97-106).

Chow, C. Y., Leong, H. V., & Chan, A. T. S. (2004). Group-based cooperative cache management for mobile clients in a mobile environment. In *Proceedings of the 33rd International Conference on Parallel Processing (ICPP)* (pp. 83-90).

Chow, C. Y., Leong, H. V., & Chan, A. T. S. (in press). Utilizing the cache space of low-activity clients a mobile cooperative caching environment. *International Journal of Wireless and Mobile Computing (IJWMC)*.

Chow, C. Y., Leong, H. V., & Chan, A. T. S, "Peer-to-Peer Cooperative Caching in a Hybrid Data Delivery Environment", In *Proceedings of the 2004 International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN2004),* 2004.

Chow, C.Y., Leong, H.V., & Chan A.T.S. (2004). Cache signatures for peer-to-peer cooperative caching in mobile environments. In *Proceedings of 2004 International Conference on Advanced Information Networking and Applications*.

Cortés, T., Girona, S., & Labarta, J. (1997). *Design issues of a cooperative cache with no coherence problems*. In *Fifth Workshop on I/O in Parallel and Distributed Systems (IOPADS'97)* (pp. 37-46).

Hara, T. (2002). Cooperative caching by mobile clients in push-based information systems. In *Proceedings of ACM International Conference on Information and Knowledge Management (ACM CIKM'02)* (pp.186-193).

Huron A.R., & Jiao, Y. (2005). Data broadcasting in mobile environment. In D. Katsaros, A. Nanopoulos, & Y. Manolopaulos (Eds.), *Wireless information highways* (Chapter 4). Hershey, PA: IRM Press.

Imielinski, T., & Badrinath, B. (1994). Mobile wireless computing: Challenges in data management. *Communications of the ACM, 37*(10), 18-28.

Imielinski, T., Viswanathan, S., & Badrinath, B. R. (1997). Data on air: Organisation and access. *IEEE Transactions on Knowledge and Data Engineering, 9*(3), 353-371.

Joseph, M. S., Kumar, M., Shen, H., & Das, S. K. (2005). Energy efficient data retrieval and caching in mobile peer-to-peer networks. In *Proceedings of 2nd Workshop on Mobile Peer-to-Peer Networks (MP2P)* (pp. 50-54).

Kara, H., & Edwards, C. (2003). A caching architecture for content delivery to mobile devices. In *Proceedings of the 29th EUROMICRO Conference: New Waves in System Architecture (EUROMICRO'03)*.

Kortuem, G., Schneider, J., Preuitt, D., Thompson, C., Fickas, S., & Segall, Z. (2001). When peer-to-peer comes face to face: Collaborative peer to peer computing in mobile ad hoc networks. In *Proceedings of the First International Conference on Peer to Peer Computing*.

Lee, W. C., & Lee, D. L. (1996). Using signature techniques for information filtering in wireless and mobile environments. *Journal on Distributed and Parallel Databases, 4*(3), 205-227.

Papadopouli, M., & Issarny, V. (2001). Effects of power conservation, wireless coverage and cooperation on data dissemination among mobile devices. In *Proceedings. of the 2nd ACM International Symposium on Mobile Ad HOC Networking and Computing (MobiHoc)* (pp. 117-127).

Prabhajara, K., Hua, K.A., & Oh, J.H. (2000). Multi-level, multi-channel air cache designs for broadcasting in a mobile environment. In *Proceedings of the 16th International Conference on Data Engineering* (pp. 167-186).

Roy, N., Roy, A., Basu, K., & Das, S. K. (2005). A cooperative learning framework for mobility-aware resource management in multi-inhabitant smart homes. In *Proceedings of 2nd Annual IEEE International Conference on Mobile and Ubiquitous Systems: Networking and Services* (pp. 393-403).

Sailhan, F., & Issarny, V. (2003). Cooperative caching in ad hoc network. In *Proceedings. of the 4th International Conference on Mobile Data Management (MDM)* (pp.13-28).

Sarkar, P., & Hartman, J. H. (2000). Hint-based cooperative caching. *ACM Transactions on Computer Systems*, *18*(4), 387-419.

Shen, H., Joseph, M. S., Kumar, M., & Das, S. K. (2005). PReCinCt: An energy efficient data retrieval scheme for ,mobile peer-to-peer networks. In *Proceedings of 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*.

Shen, H., Joseph, M. S., Kumar, M., & Das, S. K. (2004). Cooperative caching with optimal radius in hybrid wireless networks. In *Proceedings of Third IFIP-TC6 Networking Conference* (LNCS 3042, pp. 841-853).

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Indexing schemes for multi channel data broadcasting in mobile databases. *International Journal of Wireless and Mobile Computing*, *1*(6).

Waluyo, A.B., Srinivasan, B., & Taniar, D. (2005). Research in mobile database query optimization and processing. *Mobile Information Systems*, 1(4).

Xu, J., Hu, Q., Lee, D. L., & Lee, W.-C. (2000). SAIU: An efficient cache replacement policy for wireless on-demand broadcasts. In *Proceedings of the 9th International Conference on Information and Knowledge Management* (pp.46-53).

Xu, J., Hu, Q., Lee, W.-C., & Lee D. L. (2004). Performance evaluation of an optimal cache replacement policy for wireless data dissemination. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, *16*(1), 125-139.

Yajima, E., Hara, T., Tsukamoto, M., & Nishio, S. (2001). Scheduling and caching strategies for correlated data in push-based information systems. *ACM SIGAPP Applied Computing Review*, *9*(1). 22-28.

Zheng, B., Xu, J., & Lee, D. L. (2002, October). Cache invalidation and replacement strategies for location-dependent data in mobile environments. *IEEE Transactions on Computers*, *51*(10), 1141-1153.

KEY TERMS

Broadcast Environment: also known as push-based sy stem where the server would broadcast a set of data to the air for a population of mobile users to tune in for their required data.

Caching: Techniques of temporarily storing frequently accessed data designed to reduce network transfers and therefore increase speed of download

Caching Management Strategy: A strategy that relates to how a client manipulates the data that has been cached in an efficient and effective way by maintaining the data items in the client's local storage.

Cooperative Caching: A type of caching strategy that not only allows mobile clients to retrieve database items from the servers, but also from the caches of their peers.

Mobile Environment: Refers to a set of database servers that may or not be collaborative with one another that disseminate data via wireless channels to multiple mobile users.

On Demand Environment: also known as a pull–based environment, which relates to techniques that enable the server to process requests that are sent from mobile users.

Peer-to-Peer: Facilitate the features of data sharing among groups of people.

CORBA on Mobile Devices

Markus Aleksy
University of Mannheim, Germany

Axel Korthaus
University of Mannheim, Germany

Martin Schader
University of Mannheim, Germany

INTRODUCTION

Since its introduction in 1991, the Common Object Request Broker Architecture (CORBA) standard, defined by the Object Management Group (OMG, 2004b), has undergone several major revisions and has spread far throughout the domain of object-oriented and distributed systems. It not only brings about independence of computer architectures, operating systems, and programming languages, but also ensures freedom of choice with respect to Object Request Broker (ORB) product vendors. The latter benefit was the result of the introduction of a globally unique object reference, the Interoperable Object Reference (IOR), and a standard transmission protocol, the Internet Inter-ORB Protocol (IIOP), in CORBA 2.0.

CORBA uses an Interface Definition Language (IDL) to specify the interfaces that objects present to clients in order to offer their services. IDL is a purely declarative language—that is, it is used to describe the data types and interfaces in terms of their attributes, operations, and exceptions, but not to define implementation algorithms for their operations. IDL forms the foundation of CORBA's programming language independence, and language-specific IDL compilers must be used to translate IDL interface definitions into concrete programming languages. Besides the language mappings defined in the OMG standard (i.e., mappings from IDL to Ada, C, C++, COBOL, Java, Lisp, PL/1, Python, and Smalltalk), there are also non-standard language mappings to programming languages like Eiffel, Objective C, and Perl, which are exclusively implemented in certain ORB products.

While CORBA has been very successful in the domain of enterprise computing, its adoption for mobile devices is obstructed by a central problem: the limited resources of such devices. If standard-compliant CORBA-based applications are to be executed on mobile devices, storage requirements, for example, represent a major bottleneck. But for all that, several research groups have made an effort over the past few years to establish the CORBA standard in the domain of mobile devices. The existing approaches can be divided into three categories:

1. approaches that are restricted in the sense that they use an implementation of the IIOP protocol only,
2. approaches that build on the minumumCORBA specification, and
3. approaches that rely on other ways to reduce the memory footprint of a CORBA implementation.

In the following sections of this chapter, these approaches will be discussed in detail.

THE PROTOCOL APPROACH

The first alternative to realize a CORBA infrastructure on mobile devices can be described as the "protocol approach." Instead of providing an implementation of the complete CORBA specification, only the IIOP protocol is implemented in this solution.

As already mentioned, the CORBA 2.0 standard for the first time introduced the definition of a protocol for the communication between different ORBs. On an abstract level, the General Inter-ORB Protocol (GIOP) was defined, which specifies a standardized transmission syntax, the common data representation (CDR), and several message formats. Among the characteristics of CDR is a complete mapping of all data types defined in IDL and the support of different byte orders. With CORBA 2.1, the set of possible GIOP message formats was extended by "Fragment" messages. The bi-directional variant of GIOP permits both the client and the server to act as the initiator of a message for all possible kinds of messages. Since GIOP is an abstract protocol, the actual communication is routed via the Internet Inter-ORB Protocol, which provides a mapping between GIOP messages and the Transmission Control Protocol/Internet Protocol (TCP/IP) layer used in the Internet. Apart from IIOP, so-called Environment-specific Inter-ORB Protocols (ESIOPs) can be used. Currently,

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

four different IIOP versions (1.0, 1.1, 1.2, and 1.3) can be encountered. To avoid interoperability problems that might occur when ORBs implementing different protocol versions need to communicate, the OMG specifies requirements as to which protocol has to be supported in which context (cf. OMG, 2004b, pp. 15-51).

In the context of mobile devices, a challenge lies in the realization of GIOP over wireless networks. Current CORBA implementations typically use IIOP to guarantee interoperability between CORBA-based applications. However, TCP/IP is not a suitable transport layer for wireless communication (Amir, Balakrishnan, Seshan, & Katz, 1995), so better alternatives like the Mobility Layer (Haahr, Cunningham, & Cahill, 1999, 2000) or WAP (Ruggaber, Schiller, & Seitz, 1999; Ruggaber & Seitz, 2000) were developed for that purpose. In order to provide a standardized solution for this aspect too, the OMG has adopted the wireless access and terminal mobility in CORBA specification (OMG, 2005). Furthermore, to account for the fact that the Bluetooth protocol has gained increasing popularity in the area of mobile devices, the OMG has issued the GIOP Tunneling Over Bluetooth specification (OMG, 2003).

The protocol approach was implemented in the context of several projects. For example, the work described by Haahr et al. (1999) is one of the first solutions belonging to that category. Moreover, BASE (Becker, Schiele, Gubbels, & Rothermel, 2003) and LegORB (Roman, Mickunas, Kon, & Campbell, 2000) are representatives of the protocol approach. Among the defining characteristics of BASE are, according to Becker et al. (2003), the uniform access to remote services and device-specific capabilities, the decoupling of the application communication model and the underlying interoperability protocols, and its dynamic extensibility supporting the whole range of devices from simple sensors to high-capacity workstations. There are two ways to generate invocations in BASE: they can be either generated by proxies representing services, or they are encoded analogous to a CORBA DII invocation by the application developers. The "micro-broker" coming with BASE only necessitates a plug-in to transport (marshal and send) an operation invocation. The return values an invocation might possibly construct may be accepted by an additional transport plug-in.

LegORB implements a microkernel-type architecture. Its core only contains components for low-level services. Application developers have to implement specific policies or simply select them suitably, whenever they are packaged with the ORB. LegORB's core consists of three customized components: the LegORB configurator, the client-side configurator, and the server-side configurator. They provide the glue necessary to put all the components together. LegORB itself is an assembly of components with different functional scopes with duties concerning network, marshaling, demarshaling, and so forth. The actual service capability as well as the size of LegORB is determined by the number and type of components that are composed in a concrete development project.

On the one hand, the protocol approach is sufficient to enable communication between different mobile as well as stationary CORBA-based applications, but on the other hand, it has considerable limitations. The first restriction pertains to a lack of source code portability. Since the presented solutions are specifically designed to address the communication aspect, this approach requires conceptual rethinking with respect to the way the mobile CORBA applications have to be developed. This not only holds for the migration of existing CORBA-based applications to mobile devices. Moreover, modifications to the "traditional" development process used for conventional CORBA-based applications are needed to meet the changed conditions in a mobile applications setting. The conventional development process usually starts with the specification of the IDL interfaces required in the application, if static operation invocations are intended, which is the normal case. These IDL definitions are then translated using an IDL compiler. Subsequently, the developers use different standardized CORBA classes and interfaces for ORB initialization or object activation. However, the protocol-based approaches often do not provide IDL support and do not necessarily allow for the use of the stipulated classes and interfaces for routine CORBA tasks, so that they often lead to considerable learning curves, even for experienced CORBA developers, in order to adapt to the changed programming conditions.

APPROACHES BASED ON THE MINIMUMCORBA STANDARD

The first dedicated OMG specification targeted on the reduction of the footprint of CORBA-based solutions was the minimumCORBA specification (OMG, 2002). In this specification, the OMG identified parts of the full CORBA standard that might be dispensable under certain circumstances like in the case of the limited resources available on mobile devices:

> *The features of CORBA omitted by this profile clearly have value in mainstream CORBA applications. However, they are provided at some cost, in terms of resources, and there is a significant class of applications for which that cost cannot be justified.*

The omissions and reductions adopted by the OMG mainly concern the following points:

- **Omission of the Dynamic Invocation Interface:** The Dynamic Invocation Interface (DII) provides functionality that enables a client application to invoke operations of objects and to receive the returned results

at runtime without the need to have knowledge about the corresponding signatures and types at compile time.
- **Omission of the Dynamic Skeleton Interface:** The Dynamic Skeleton Interface (DSI) became part of the CORBA standard in version 2.0. It represents a runtime mechanism to integrate components that do not have any IDL-based compiled skeletons at compile time. The DSI provides an interface that is able to accept a specific type of invocation. Incoming messages are analyzed with regard to their intended receiver object and the operation to be performed. The DSI can process both static and dynamic operation invocations.
- **Omission of the Interface Repository:** The interface repository (IR) is a runtime repository containing machine-readable versions of IDL definitions. It provides the type information needed by the DII. With CORBA 2.0, globally unique identifiers for components and their interfaces were introduced, which are system generated, so that the ORB is able to deal with entries from different IRs without the risk of ambiguities.
- **Omission of the DynAny Functionality:** The functionality of this interface allows for runtime generation of new data types that are unknown at compile time.
- **Reduction of the ORB Interface:** The operations omitted from the ORB interface specification provided functionality that was required for DII operation and could therefore be disposed of.
- **Reduction of the Object Interface:** The operations omitted from the Object interface specification mainly provided functionality that was required for DII and IR operation, were relevant for older CORBA versions, or required types that are omitted in the profile.
- **Reduction of the Portable Object Adapter Interface:** By introducing the Portable Object Adapter (POA) interface as replacement of the underspecified Basic Object Adapter (BOA), portability of CORBA-based applications could be increased. The POA serves as a link between the ORB and the servants. In the context of the minimumCORBA specification, its functionality has been reduced.

An example of an approach that belongs to the category of minimumCORBA-based solutions is K-ORB, which was created by the Distributed Systems Group in Dublin (cf. http://www.dsg.cs.tcd.ie).

A central advantage of approaches relying on minimumCORBA is the fact that the application development process remains largely unchanged. Although the use of well-known classes and interfaces like ORB and POA is restricted, it is still possible. This approach therefore requires considerably less reorientation on the part of developers already familiar with the construction of CORBA-based applications. Relying on this approach, it is much easier to port existing applications or to realize new projects than with the protocol approach.

OTHER APPROACHES

Other endeavors of research groups to achieve a reduction of the scale and scope of CORBA-based applications for their deployment on mobile devices can be summarized as being based on the following general ideas:

- using a microkernel architecture,
- following a component-based approach,
- employing a reflection-based technique, or
- providing capabilities for reconfiguration and adaptation.

The projects *MICO on Palm Pilot* and *Mico on iPAQ* (Puder, 2002) were among the first attempts to establish MICO ORB (Puder, 2000), an open source implementation of CORBA, in the domain of mobile applications. They used MICO's microkernel architecture to reduce the required memory. The use of a microkernel had already been tested successfully in several architectures and was generalized in the description of the microkernel design pattern in Buschmann, Meunier, Rohnert, Sommerlad, & Stal (1996). The developers of ZEN ORB start from the basic premise that the simultaneous support of all protocol versions at the same time is not generally required. Thus, the introduction of a "plug-and-play" interface, which can dynamically reload components on demand—such as protocol classes of a specific version—is seen as a possible solution to reduce memory requirements. Klefstad, Rao, and Schmidt (2003) describe how such an ORB could be designed and identify additional optimization factors, such as object adapters, protocol transports, object resolvers, IOR parsers, any data type handlers, buffer allocators, and CDR streams.

Most of the documented approaches, however, rely on a mix of concepts. K-ORB and ZEN, for example, use component technology together with reconfiguration capabilities. LegORB and Universally Interoperable Core (Roman, Kon, & Campbell, 2001) implement all of the architectural concepts discussed so far.

USING CORBASERVICES ON MOBILE DEVICES

The ORB represents the fundamental communication component in distributed CORBA applications. However, to support the application developer further, the OMG has standardized a number of system-level services, called CORBAservices. These services extend the basic functionality provided by the ORB by offering frequently required functionality (e.g., concerning service lookup, event management, transactions, etc.).

Although the use of CORBAservices promotes aspects like portability and reusability and enables experienced developers to produce their applications more quickly, the back side is that they generate additional memory requirements. Even though it is not necessary to run the various services directly on the mobile devices (they are normally hosted on PCs or workstations), mobile applications that need to make use of those services still need to have the corresponding stub files generated by an IDL compiler at their disposal locally, that is, on the mobile device. Typically, a CORBA-based application does not require all of the CORBAservices available. On the contrary, most applications do without CORBAservices completely, employ only the Naming Service, or use a relatively small amount of additional services, depending on their application purpose.

For PDAs, which often do not have any storage facilities apart from their Random Access Memory (RAM), stubs of CORBAservices that are never needed by a client application running on that PDA would undesirably lock memory resources that are urgently needed for other purposes.

One way to reduce the amount of memory required by those stub files is to use OMG's Lightweight Services specification (OMG, 2004a). It only contains the definition of three functionally reduced CORBAservices: the Lightweight Naming service, the Lightweight Event service, and the Lightweight Time service. In Aleksy, Korthaus, and Schader (2003), we use the so-called "Janus" approach to make the memory footprint significantly smaller. The basic idea of the Janus approach is to adapt the appearance of a component to its specific use. In our case, we have simplified and reduced the management interfaces of the event service as perceived by the mobile application. To this end, we developed a dedicated component, called Event Service Proxy, which offers a drastically simplified version of the event service interfaces and is responsible for the main part of the registration process. The actual event transmission occurs as always—that is, without utilizing the Event Service Proxy. In this way, the approach does not affect the performance of the application, except for the registration process.

FUTURE TRENDS

An important challenge for the future is the extension of the spectrum of lightweight CORBAservices to be used by mobile applications. For example, the design and realization of a "mobile" variant of the transaction service (OMG, 2000b) will be of crucial importance for the advancement of m-business applications. Similarly, a "mobile" version of the trading service (OMG, 2000a) is needed in order to provide CORBA applications running on mobile devices with extended functionality for the detection of new services.

CONCLUSION

In this article, we have presented different research efforts aiming at the goal of establishing CORBA in the domain of mobile devices. While the protocol-based approach is easier to implement and enjoys great popularity despite its limitations, the minimumCORBA-based approach does not yet meet the expectations. The ongoing advancement of mobile devices and corresponding communication infrastructures, however, raise hope that this approach will increasingly attract attention in the near future. Another future challenge will be the use of CORBAservices on mobile devices. By adopting the Lightweight Services specification, the OMG has made a first step in that direction. Nevertheless, this aspect will continue to necessitate the development of ideas for new solutions in the years to come.

REFERENCES

Aleksy, M., Korthaus, A., & Schader, M. (2003). CARLA—a CORBA-based architecture for lightweight agents. *Proceedings of the International Conference on Intelligent Agent Technology* (IAT'03) (pp. 111-118), Halifax, Canada.

Amir, E., Balakrishnan, H., Seshan, S., & Katz, R.H. (1995). Efficient TCP over networks with wireless links. *Proceedings of the 5th IEEE Workshop of Hot Topics in Operating Systems,* Orcas Island, WA.

Becker, C., Schiele, G., Gubbels, H., & Rothermel, K. (2003). BASE—a micro-broker-based middleware for pervasive computing. *Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications* (PerCom'03), Fort Worth, TX.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., & Stal, M. (1996). *Pattern-oriented software architecture—a system of patterns.* Chichester: John Wiley & Sons.

Haahr, M., Cunningham R., & Cahill, V. (1999). Supporting CORBA applications in a mobile environment. *Proceedings of the 5th International Conference on Mobile Computing and Networking* (MobiCom'99), Seattle, WA.

Haahr, M., Cunningham, R., & Cahill, V. (2000). Towards a generic architecture for mobile object-oriented applications. *Proceedings of the Workshop on Service Portability and Virtual Customer Environments (SerP 2000)*, San Francisco, CA.

Klefstad, R., Rao, S., & Schmidt, D. (2003). Design and performance of a dynamically configurable messaging protocols framework for real-time CORBA. *Proceedings of the 36th Hawaii International Conference on System Sciences* (HICSS-36), Big Island, HI.

OMG (Object Management Group). (2003). *GIOP tunneling over Bluetooth.* OMG Technical Document Number dtc/2003-10-03. Retrieved from http://www.omg.org/cgi-bin/doc?dtc/2003-10-03

OMG. (2004a). *Lightweight services specification, version 1.0.* OMG Technical Document Number formal/04-10-01. Retrieved from ftp://ftp.omg.org/pub/docs/formal/04-10-01.pdf

OMG. (2002). *Minimum CORBA specification.* OMG Technical Document formal/02-08-01. Retrieved from http://www.omg.org/cgi-bin/ doc?formal/2002-08-01.pdf

OMG. (2004b). *The Common Object Request Broker: Architecture and specification. Version 3.0.3.* OMG Technical Document Number formal/04-03-01. Retrieved from ftp://ftp.omg.org/pub/docs/formal/04-03-01.pdf

OMG. (2000a). *Trading object service specification. Version 1.0.* OMG Technical Document Number formal/00-06-27. Retrieved from http://www.omg.org/cgi-bin/doc?formal/2000-06-27

OMG. (2000b). *Transaction service specification. Version 1.4.* OMG Technical Document Number formal/03-09-02. Retrieved from http://www.omg.org/cgi-bin/doc?formal/2003-09-02

OMG. (2005). *Wireless access and terminal mobility in CORBA. Version 1.2.* OMG Technical Document Number formal/05-05-02. Retrieved from http://www.omg.org/cgi-bin/doc?formal/2005-05-02

Puder, A. (2002). Middleware for handheld devices. *Proceedings of the AT&T Software Symposium,* Middletown, NJ.

Puder, A., & Römer, K. (2000). *MICO—an open source CORBA implementation.* San Francisco: Morgan Kaufmann.

Roman, M., Mickunas, D., Kon, F., & Campbell, R.H. (2000). LegORB and ubiquitous CORBA. *Proceedings of the Workshop on Reflective Middleware (IFIP/ACM Middlewar 2000)*, IBM Palisades Executive Conference Center, NY.

Roman, M., Kon, F., & Campbell, R.H. (2001). Reflective middleware: From your desk to your hand. *Distributed Systems Online, 2*(5).

Ruggaber, R., Schiller, J., & Seitz, J. (1999). Using WAP as the enabling technology for CORBA in mobile and wireless environments. *Proceedings of the 7th IEEE Workshop on Future Trends of Distributed Computing Systems* (pp. 69-74), Cape Town, South Africa.

Ruggaber, R., & Seitz, J. (2000). Using CORBA applications in nomadic environments. *Proceedings of the 3rd IEEE Workshop on Mobile Computing Systems and Applications* (pp. 161-170), Monterey, CA.

KEY TERMS

CORBAservices (a.k.a. CORBA Services): General-purpose, system-related extensions of the core functionality of an ORB that are relevant to the basic operation of a distributed application.

General Inter-ORB Protocol (GIOP): Protocol specifying a standardized transmission syntax, together with several message formats.

Interface Definition Language (IDL): A purely declarative language used for the description of a CORBA application's data types and interfaces—with their attributes, operation signatures, and exceptions—independently of a concrete implementation language.

Internet Inter-ORB Protocol (IIOP): Protocol specifying how GIOP messages can be exchanged over Transmission control protocol/Internet protocol (TCP/IP) connections.

Object Adapter: Technically, the connecting link between the ORB and the proper object implementations. From a logical perspective, it connects CORBA objects that are specified by means of IDL to their implementations, which were written in a concrete programming language.

Object Request Broker (ORB): Constitutes the architecture's communication component and sometimes denoted as the "object bus."

Reflection: Means that a program is able to gain insight into its own structure. Reflection makes it possible to query meta-information about classes and their instances at runtime. Also called Introspection.

Cross-Layer RRM in Wireless Data Networks

Amoakoh Gyasi-Agyei
Central Queensland University, Australia

INTRODUCTION

A wireless data network is the infrastructure for mobile computing, which is the act of communicating while on-the-move via portable computers. Hence, wireless data networks (WDNs) and associated issues are enabling technologies for mobile computing. Usually, a WDN does not stand alone; it is connected to the fixed Internet. Hence, it is also referred to as *wireless Internet.* The unit of information transfer and processing in a packet-switched WDN is packet, which is a bunch of bits with the identifying fields needed for efficient forwarding. Advances in digitization enable a packet's content to be of varying nature, such as conversational voice samples, streaming video scenes, or non-real-time data.

Every WDN allowing multiple users to share its services requires a radio resource management (RRM) to coordinate the efficient use of the wireless transmission medium or channel. The wireless channel used by WDNs has time-, frequency-, and environment-dependent quality due to multi-path signal propagation, shadowing, path loss, and user mobility. There are two main RRM philosophies: link/rate adaptation, and opportunistic communications or opportunistic RRM. The link adaptation philosophy views the inherent channel variability as bad and hence mitigates it via complex mechanisms such as interleaving, power control, equalization, and spatial diversity (Gyasi-Agyei, 2005).

Opportunistic communications is the recent RRM philosophy, which exploits the dynamics in wireless channel quality to improve system throughput. In fact, measures have been proposed to enhance or even induce channel variability if necessary to further improve throughput (Viswanath, Tse, & Laroia, 2002). RRM has several functionalities, such as traffic admission control, power control, and scheduling. This chapter focuses on opportunistic communications or scheduling (OS). In fact, link/rate adaptation-based RRM is also somewhat cross-layer protocol engineering, as its uses physical layer information, but for a different purpose than OS. In principle, OS can be designed for single-carrier or multi-carrier, single-antenna or multi-antenna, single-hop or multi-hop, centralized or distributed wireless networks. It can also serve both real-time and non-real-time network traffic. However, it is much easier to design an OS scheme for single-hop, centralized wireless networks serving non-real-time data traffic. The achievable throughput gains are also maximized when no traffic timing constraints are embedded in the OS policy.

Scheduling is the dynamic process of allocating a shared resource to multiple parallel users in order to optimize some desirable performance metrics. Metrics of interest include: maximization of system throughput, minimization of packet delay and jitter, and the provision of fairness. Scheduling is a key mechanism in RRM and operates in the medium access control (MAC) layer. Only three things can happen to the transmission medium of a multiuser network: resource hogging, resource clogging, or equitable resource sharing. Without a MAC protocol, the desirable third option can hardly occur. In the following we discuss some general aspects of OS, propose a generalized OS design framework, discuss future trends of OS, and list some open issues in OS design.

BACKGROUND

This section reviews some background material on OS.

Why Cross-Layer Protocol Engineering?

A protocol is a set of agreed-upon rules by which two entities communicate efficiently. This includes both semantics and syntax. The telecommunications industry has been sustained so far by the open systems interconnect (OSI) reference architecture designed by the International Standards Organization (ISO). Hence this architecture is referred to as the ISO/OSI model. The ISO/OSI model is a brick-wall protocol architecture whereby the functions of a complete workable communications system is broken into a set of functionalities (not without duplications) and each set is called a layer. Each layer provides a service to the overlying layer. However, the service user is not privy to how the service is provided, and neither does it know the features of its service provider. Non-adjacent layers on the same machine have no interaction, but layers at the same level on two machines communicating interact logically to exchange data and signaling. This horizontal interaction is referred to as *peer-to-peer communication.* Indeed, no practical system is strictly designed according to the ISO/OSI model. However, the popular protocol used on the Internet, TCP/IP model, is designed based on the ISO/OSI philosophy.

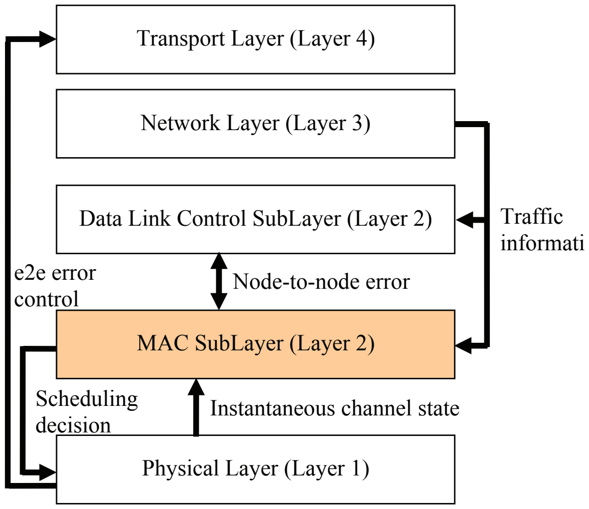
The ISO/OSI philosophy has been embraced so far as great, as its modularity enables upgrading of one layer with minimal impact on other layers, until the recent drive towards system performance optimization via cross-layer engineering.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Protocol layering and modularity vs. protocol efficiency

Layering Protocol Engineering	Cross-Layer Protocol Engineering
Traditional protocol design approach	Modern protocol design approach
Modularity oriented	Efficiency oriented
Strict layering abstraction, no interactions between non-adjacent protocol layers on the same machine	Exploits inter-layer interactions to optimize overall system performance
Partitioning of protocol functionalities reduces protocol complexity	High computational complexity
Allows protocol specialization	Requires interdisciplinary knowledge
	Incompatible with existing protocols
Easy to manage and maintain	Complicated system management and maintenance
	Revolutionary approach to system design
Only lower protocol layers are network specific	Entire protocol may be network specific
	Can reduce device energy consumption

Figure 1. An OS algorithm showing cross-layer signaling exchanges



This new era of *cross-layer protocol engineering* has opened both opportunities and challenges for protocol designers. Fourth-generation networks and beyond are expected to exploit the benefits in cross-layer protocol engineering (Shakkottai, Rappaport, & Karlsson, 2003). *Cross-layer protocol design leverages runtime information across different layers to enhance the performance of the entire wireless system.* It can also be used to reduce functionality duplications during protocol design. The idea originates from the notion that layers can adapt to the instantaneous condition of other layers to improve the overall service provided to network applications. Table 1 summarizes the basic features of both protocol design philosophies.

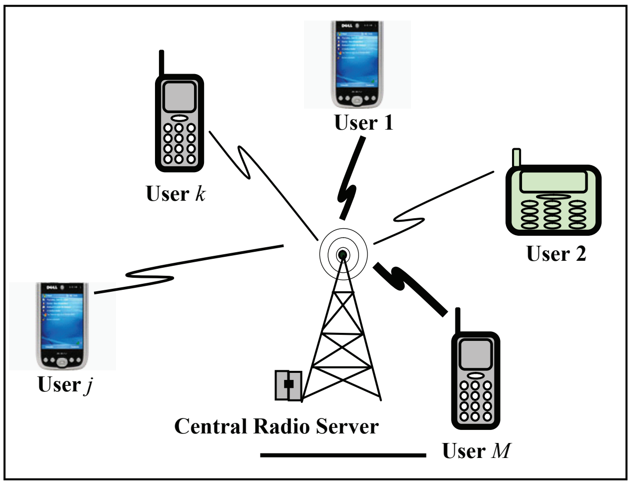
Figure 1 illustrates the design of a cross-layer wireless scheduler. We can observe interactions in several directions between layers 1-4. These inter-layer interactions can allow higher layers to adapt to the random variability in physical layer properties to optimize system performance. Specifically, Figure 1 shows channel state information (CSI) from Layer 1 and traffic information from Layer 3 communicated to Layer 2, where the scheduler operates. These parameters enable the MAC layer to make informative decisions. For example, the

CSI can be in the form of instantaneous channel signal-to-noise ratio, the received signal strength, mobile user information, user speed of motion, or channel impulse response (Verikouskis, Alonso, & Giamalis). The traffic information can be the service history, required packet timing constraints, minimum throughput requirement, or minimum reliability requirements. The traffic delay information and the CSI enable the link layer to use the appropriate error control scheme depending on channel error pattern and traffic delay requirements. That is, whether backward-error or forward-error, correction should be used. The delay information also influences the scheduling decision. The CSI feedback to the connection-oriented transport layer allows the latter's end-to-end (e2e) error control mechanism to distinguish congestion-related from channel-state-related packet losses.

Why Opportunistic Scheduling?

As discussed above, the wireless channel quality varies with time, space, and operating frequency—that is, specto-spatio-temporally varying. Hence, multiple users sharing a wireless resource experience independently varying channel qualities at the same time: some users experience poor channels, while others experience good channels. This is referred to as *multiuser diversity* (Knopp & Humblet, 1995), the basis of opportunistic scheduling. The origin of opportunistic communications is attributed to the works of Knopp and Humblet (1995) and Viswanath et al. (2002). Opportunistic scheduling (Shakkottai & Stolyar, 2001; Gyasi-Agyei, 2003; Liu, Chong, & Shroff, 2003; Liu, Grul, & Knightly, 2003; Hu, Zhang, & Sadowsky, 2004; Gyasi-Agyei & Kim, 2006) is a wireless scheduling which exploits multiuser diversity to maximize the total system throughput. The per-user throughput is also maximized if users have symmetric fading statistics. The throughput gain is achieved by picking a user among all active users at a given time, which has the relatively best channel condition and hence the highest data transfer speed. The achieved gain in OS over a comparable non-OS scheme is referred to as *multiuser diversity gain*. Multiuser diversity gain increases with the rate of channel variations, the dynamic range of channel variability and randomization, and hence the size of user population sharing a wireless resource. The larger the user population, the higher the chance of picking a user at a high channel quality at any time, and hence the higher the multiuser diversity gain. This is the motivation behind proposals to enhance channel scattering if necessary (Viswanath et al., 2002). Besides the above limitations, the wireless spectrum is a finite resource. Hence, novel efforts such as opportunistic communications are required to maximize its utility. While conventional diversity techniques combat channel slow fading to achieve error-free communications, multiuser diversity exploits channel fast fading to boost system throughput.

Figure 2. Time-slotted, single-cell system using OS to serve multiple heterogeneous users



Consider Figure 2 where M users share a wireless server offering time-slotted service. The feasible data transmission speed of each user varies over time in accordance with its channel quality variations. The maximum speed that user m can send data per Hertz of bandwidth at time t lies in the range:

$$0 \leq R_m(t) < \min\left[\log_2\left(1+SNR_m(t)\right),\ 2\log_2 K\right] \tag{1}$$

where $K \geq 1$ is the number of signaling levels used in the system and $SNR_m(t)$ is user m's instantaneous signal-to-noise ratio. By scheduling a user with the highest instantaneous rate, a greedy OS is able to achieve the throughput bound at the cost of all other system performance metrics. However, guarantee on bounded delay is necessary if OS has to support real-time traffic. This can be achieved by embedding traffic temporal parameters into OS and making optimum trade-off between delay, throughput, and fairness.

Design Issues and Performance Metrics of Cross-Layer RRM Algorithms

The design of efficient cross-layer RRM mechanisms requires the consideration of several issues. Some of the crucial design metrics are: throughput, energy efficiency, equity, algorithmic complexity, scalability and optimality, feasibility, stability, timeliness, algorithmic convergence, near-far, hidden-terminal, and exposed-terminal problems.

The wireless medium used by WDNs is a finite resource which is getting crowded, requiring an economized usage. Energy efficiency is an interesting issue due to the power constraints of battery-powered wireless terminals. Also, in some network architectures, notably wireless sensor networks operating in a harsh environment, the power of the wireless nodes cannot be easily replenished. Hence, power failure of a single node can destabilize an entire network. RRM schemes must be fair to prevent hogging and/or starvation. However, fairness does not necessary mean that all queues receive the same level of service. For this reason several fairness models are designed; examples are utilitarian fairness and proportional fairness.

Algorithms with low complexity reduce energy dissipation in wireless terminals and prolong their battery lifetime. Low-complexity algorithms can also reduce the processing time of traffic and hence reduce the end-to-end packet delay. Scalable and stable algorithms are able to support network growth under all traffic patterns. Hence, RRM algorithms should answer questions like: Can the algorithm work efficiently in networks of all sizes? Is the algorithm stable under all possible traffic arrival patterns and user population? Stability ensures that the length of every queue in an *admissible system* is bounded at steady state—that is:

$$\lim_{x\to\infty}\sup E\left[\left\|\mathbf{q}_t\right\|\right]\leq\alpha$$

where

$$\mathbf{q}_t=(q_t^1,q_{t,}^2\cdots q_t^n)\qquad(2)$$

where q_t^k is the length of the *k*th queue at time *t*, *n* is the number of simultaneous queues in the system at a given time, and α is a small non-negative number. Algorithm convergence is necessary to maintain a stable system. A queuing system is said to be admissible if there is a service process which is able to maintain stable queues under a given traffic arrival process.

The near-far problem (Figure 3c) occurs when two transmitters send signals to the same receiver about the same time and one has a much stronger signal than the other at the receiver. Thus the stronger signal prevents the receiver from detecting the weaker signal. This term is coined, as the transmitter of the stronger signal is usually closer to the receiver than that of the weaker signal. This issue is more troublesome in CDMA systems. One popular solution to this problem is the use of transmitter power control.

Figure 3. Illustration of some wireless communication issues

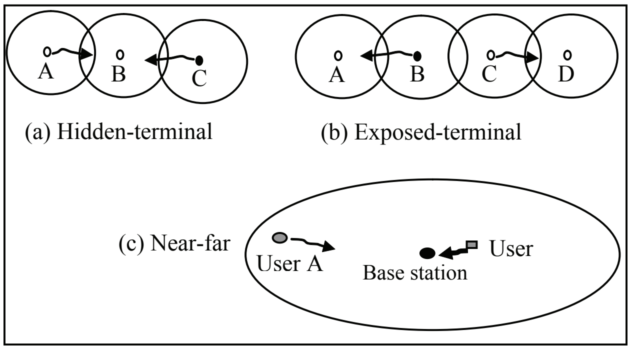


Figure 3 illustrates the *hidden-terminal* and the *exposed-terminal* problems in wireless networks. The former occurs when two nodes (here A and C) that are out of range with each other transmit signals to the same node (here B) and collide at the receiver B. The exposed-terminal problem occurs when an ongoing communication between two users unnecessarily prevents communication between another pair of users. Combining power control with scheduling prevents all the issues in Figure 3.

A Framework for Utility-Based Opportunistic Schedulers

Utility functions are widely applied in economics to quantify the benefit in using finite resources. Maximizing a utility function results in maximizing the benefits in the corresponding resources. In terms of wireless communications, the resources are power, transmission channels, bandwidth, and so forth. Let *x* be a vector of network parameters or their indices. We can thus define parametric utility functions *U*(*x*,*t*) at any time *t* which consider the RRM design issues discussed above which are of interest in a given situation. Such a generalized parametric model enables the classification of several OS schemes in the literature. Consider the three-part utility function:

$$U(x,t)=s(x,t)\cdot f(x,t)\cdot d(x,t)\qquad(3)$$

For example, maximizing *f*(*x*,*t*) means enhancing fairness and maximizing *s*(*x*,*t*) enhances throughput. *U*(*x*,*t*) in (3) is simple and covers three key OS design metrics.

As a simple design example, consider the time-slotted multiuser system shown in Figure 2, and let $s(x,t) = R_m(t)$, $f(x,t) = 1/B_i(t)$, and $d(x, t) = \exp[\{\kappa_x W_x(t) - \eta(t)\}/q_x(t)]$.

This results in the delay-aware BLOT (D-BLOT) policy, a variation of the Best Link LOwest Throughput (BLOT) first scheduling (Gyasi-Agyei & Kim, 2006). The resulting utility function is:

$$U(x,t)=\frac{R_m(t)}{B_i(t)}\exp\left[\frac{\kappa_x W_x(t)-\eta(t)}{q_x(t)}\right]$$

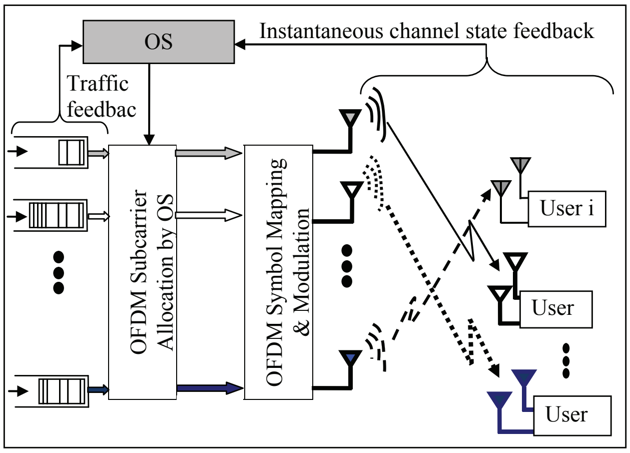
Here, $x = (B_i(t), R_m(t))^T$ or $x = (i,m)^T$.For simplicity we can require that $\eta(t) = \sum_x(t)/\|x\|$ and $q_x(t) = \eta(t)$. $W_x(t)$ is the waiting time of connection *x* at time *t*, and κ_x is a constant factor reflecting a connection's timing constraints. We note that any of the three functionals in equation (3) can be set to unity. For example, all non-real-time OS schemes use $d(x,t) = 1$. It is worthy to remark that both BLOT and D-BLOT guarantee a minimum service to multiple queues that are concurrently active at a given user, as detailed in Gyasi-Agyei and Kim (2006) and Gyasi-Agyei (2005).

FUTURE TRENDS AND OPEN ISSUES

Gyasi-Agyei and Kim (2006) classify OS schemes into eight types under four subgroups and recognize that Type F OS is a field yet to be explored. Li and Niu (2004) attempt this problem. Below, we discuss some possible scenarios for such OS. Figure 4 illustrates the downlink of a multi-carrier (OFDM) wireless system using multiple antennas, OS, and dynamic subcarrier assignment. In this architecture, each OFDM symbol carries the data for a single connection, and disjointed subsets of OFDM subcarriers are allocated to different connections in a given time slot. This provides mutual orthogonality across the signals received at the receiver, as the spatial signatures are usually correlated. Elements of the antenna array at the transmitter transmit data of independent connections. However, multiple active connections may terminate at the same receiver. Assume that there are X concurrent connections and L multiple antennas per radio server. If $L=X$, then X mutually orthogonal sets of random beams can serve the X connections simultaneously. If $L>X$, then spatial multiplexing gain can be exploited on the remaining $L-X$ beams.

The architecture in Figure 4, whose details are left for future research, may operate as follows. Channel qualities of OFDM subcarriers from Layer 1 and traffic information from the application layer through Layer 3 are fed into the OS. The OS then uses this information to build a utility function. In each scheduling epoch, the queue that has the highest utility on a subcarrier is allowed to transmit/receive information on that subcarrier. The set of subcarriers allocated to each connection in each time slot is used by the OFDM transmitter to construct an OFDM modulating symbol and then transmitted to the corresponding user.

Figure 4. A downlink architecture for Type F opportunistic scheduler


OS
Instantaneous channel state feedback
Traffic feedbac
OFDM Subcarrier Allocation by OS
OFDM Symbol Mapping & Modulation
User i
User
User

OFDM enables broadband communications over otherwise frequency-selective fading channels without remarkable inter-symbol interference. Time-slotted OFDM-based networks exploit the synergies between frequency and time multiplexing. Another advantage in OFDM transmissions is that different power levels can be allocated to different subcarriers (i.e., adaptive subcarrier allocation) based on their runtime qualities to meet a given reliability requirement. This is referred to as the multiuser *water-filling principle.* One issue with this approach is the signaling load involved in estimating the CSI on all subcarriers and communicating them to the scheduling engine. A subcarrier clustering approach bundling a set of subcarriers into a subchannel and assigning a single CSI has been proposed to reduce this issue.

Mobile computing and its enabling infrastructure are faced with several challenges. For example, as user terminals must be carried around, they impose ergonomic constraints. These constraints in turn penalize power supply, device memory size, disk capacity, and processor speeds. Although these issues have been improved in recent years, more remains to be done to boost the technology's uptake. These end-user terminal constraints affect the issues discussed under 'Design Issues of RRM Algorithms'. Some of the open issues and challenges in OS include the need to:

- Develop efficient mechanisms using CSI to help a connection-oriented transport layer to distinguish channel-dependent packet losses from network congestion-dependent packet losses so as to minimize throughput degradation due to end-to-end congestion control mechanisms. Current proposals to solve this problem include the use of the explicit congestion notification flag in packet header, snoop transport protocol, and explicit loss notification (Jiang, Zhuang, & Shen, 2005).
- Develop channel-state-dependent data compression schemes to reduce data transfer times.
- Investigate implications of cross-layer design on the overall networking architecture.
- Develop CSI feedback without errors and with minimal delay. Also, the effect of errors and delays in CSI on scheduling performance is an interesting task. Estimation and communication of CSI from physical to higher layers can be quite signaling intensive and hence impact transmission efficiency. This issue is exacerbated in networks with high mobility and/or changing topology. Hence, novel techniques with optimum trade-off between currency of CSI and frequency of CSI estimation are needed.

- Develop OS for multi-hop and distributed wireless networks, for example, sensor networks.
- Design OS schemes to serve traffic of varying characteristics over varying interfaces.
- Develop OS for multi-carrier wireless networks using spatial diversity.
- Design OS for variable-size time slots to serve variable-length packets.
- Develop simple but efficient online OS algorithms for wireless networks.

CONCLUSION

We have presented a handy introduction to opportunistic communications, an idea whose time has come. It is hoped that the open issues underscored become a basis for further R&D on the topic. Opportunistic communications is inherently cross-layer in nature and a disruptive technology, as it makes fading-combating techniques counter-productive. This is the reality facing the well-established RRM based on ISO/OSI model and link adaptation.

REFERENCES

Gyasi-Agyei, A. (2003, September). BL^{2}xF–channel state-dependent scheduling algorithms for wireless IP networks. *Proceedings of the IEEE International Conference on Networks* (pp. 623-628). Sydney.

Gyasi-Agyei, A. (2005). Multiuser diversity based opportunistic scheduling for wireless data networks. *IEEE Communications Letters*, *9*(7), 670-672.

Gyasi-Agyei, A., & Kim, S.-L. (2006, January). Comparison of opportunistic scheduling policies in time-slotted AMC wireless networks. *Proceedings of the IEEE International Symposium on Wireless Pervasive Computing* (pp. 1-6). Phuket, Thailand.

Gyasi-Agyei, A., & Kim, S.-L. (2006). Cross-layer multiservice opportunistic scheduling for wireless networks. *IEEE Communications Magazine, 44*(6).

Hu, M., Zhang, J., & Sadowsky, J. (2004). Traffic aided opportunistic scheduling for wireless networks: Algorithms and performance bounds. *Elsevier Computer Network,* (November), 505-518.

Jian, H., Zhuang, W., & Shen, H. S. (2005). Cross-layer design for resource allocation in 3G wireless networks and beyond. *IEEE Communications Magazine,* (December), 120-126.

Knopp, R., & Humblet, P.A. (1995, June). Information capacity and power control in single-cell multiuser communications. *Proceedings of the IEEE International Conference on Communications* (pp. 331-335). Seattle, WA.

Li, L., & Niu, Z. (2004, September). A multi-dimensional radio resource scheduling scheme for MIMO-OFDM systems with channel dependent parallel weighted fair queuing (CD-PWFQ). *Proceedings of the IEEE International Symposium on PIMRC* (pp. 2367-2371).

Liu, X., Chong, E. K. P., & Shroff, N. B. (2003, October). A framework for opportunistic scheduling in wireless networks. *Computer Networks, 41*(4), 451-474.

Liu, Y., Grul, S., & Knightly, E. W. (2003). WVFQ: An opportunistic wireless scheduler with statistical fairness bounds. *IEEE Transactions on Wireless Communications, 2*(5), 1017-1028.

Shakkottai, S., Rappaport, T. S., & Karlsson, P. C. (2003). Cross-layer design for wireless networks. *IEEE Communications Magazine, 41*(10), 74-80.

Verikouskis, C., Alonso, L., & Giamalis, T. (2005). Cross-layer optimization for wireless systems: A European research key challenge. *IEEE Communications Magazine, 43*(7).

Viswanath, P., Tse, D. N. C., & Laroia, R. (2002). Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory, 48*(6), 1277-1294.

KEY TERMS

Cross-Layer Protocol (CLP): Any communications protocol that interacts with a protocol operating on any other layer of the ISO/OSI protocol model.

Fading: Dynamically changing attenuation, usually experienced on wireless channels.

Mobile Computing: Communication via portable computer over a wireless data network while on-the-move.

Multiuser Diversity (MUD): The statistically independent variability of channel qualities (states) across multiple users in a multiuser system at a given time.

Multiuser Diversity Gain (MDG): The gain achieved (compared to a non-channel-aware version of the same algorithm) when a radio resource manager (RRM), such as a packet scheduler, exploits MUD on wireless connectivity.

Opportunistic MAC (OMAC): A medium access control protocol that exploits MUD in its operation. An example of OMAC is an opportunistic scheduling. OMAC is also referred to as opportunistic communications.

Opportunistic Scheduler (OS): A traffic scheduler that utilizes MUD to enhance desirable system performance such as system throughput or spectral efficiency. OS is also referred to as channel-aware or channel-state-dependent scheduling.

Radio Resource Management: The process of allocating wireless network resources (e.g., transmission channels, power, and spectrum) to radio nodes in an optimum manner—optimum in the sense that the service requirements of most of the users are met and system throughput is maximized.

Scheduling: The dynamic process of allocating a shared resource to multiple competing users in order to optimize some desirable performance metrics.

Data Caching in Mobile Ad-Hoc Networks

Narottam Chand
Indian Institute of Technology Roorkee, India

R. C. Joshi
Indian Institute of Technology Roorkee, India

Manoj Misra
Indian Institute of Technology Roorkee, India

INTRODUCTION

Mobile wireless networks allow a more flexible communication structure than traditional networks. Wireless communication enables information transfer among a network of disconnected, and often mobile, users. Popular wireless networks such as mobile phone networks and wireless local area networks (LANs), are traditionally infrastructure based—that is, base stations (BSs), access points (APs), and servers are deployed before the network can be used. A mobile ad hoc network (MANET) consists of a group of mobile hosts that may communicate with each other without fixed wireless infrastructure. In contrast to conventional cellular systems, there is no master-slave relationship between nodes, such as base station to mobile users in ad-hoc networks. Communication between nodes can be supported by direct connection or multi-hop relays. The nodes have the responsibility of self-organizing so that the network is robust to the variations in network topology due to node mobility as well as the fluctuations of the signal quality in the wireless environment. All of these guarantee anywhere and anytime communication. Recently, mobile ad-hoc networks have been receiving increasing attention in both commercial and military applications.

The dynamic and self-organizing nature of ad-hoc networks makes them particularly useful in situations where rapid network deployments are required or it is prohibitively costly to deploy and manage network infrastructure. Some example applications include:

- attendees in a conference room sharing documents and other information via their laptops and PDAs (personal digital assistants);
- armed forces creating a tactical network in unfamiliar territory for communications and distribution of situational awareness information;
- small sensor devices located in animals and other strategic locations that collectively monitor habitats and environmental conditions; and
- emergency services communicating in a disaster area and sharing video updates of specific locations among workers in the field, and back to headquarters.

Unfortunately, the ad-hoc nature that makes these networks attractive also introduces many complex communication problems. From a communications perspective, the main characteristics of ad-hoc networks include:

1. lack of pre-configuration, meaning network configuration and management must be automatic and dynamic;
2. node mobility, resulting in constantly changing network topologies;
3. multi-hop routing;
4. resource-limited devices, for example, laptops, PDAs, and mobile phones have power and CPU processing constraints;
5. resource-limited wireless communications, for example, a few kilobits per second per node; and
6. potentially large networks, for example, a network of sensors may comprise thousands or even tens of thousands of mobile nodes.

A key research challenge in ad-hoc networks is to increase the efficiency of data transfer, while handling the harsh environmental conditions such as energy constraints and highly mobile devices. Presently, most of the researches in ad-hoc networks focus on the development of dynamic routing protocols that can improve the connectivity among mobile nodes which are connected to each other by one-hop/multi-hop links. Although routing is an important issue in ad-hoc networks, other issues such as information/data access are also very important since the ultimate goal of using such networks is to provide information access to mobile nodes. One of the most attractive techniques that improves data availability is caching. In general, caching results in:

1. enhanced QoS at the nodes—lower jitter, latency, and packet loss;

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

2. reduced network bandwidth consumption; and
3. reduced data server/source workload.

In addition, reduction in bandwidth consumption infers that a properly implemented caching architecture for ad-hoc network can potentially improve battery life in mobile nodes.

BACKGROUND

Caching has been proved to be an important technique for improving the data retrieval performance in mobile environments (Chand, Joshi, & Misra, 2004, 2005; Cao, 2002, 2003). With caching, the data access delay is reduced since data access requests can be served from the local cache, thereby obviating the need for data transmission over the scarce wireless links. However, caching techniques used in one-hop mobile environments (i.e., cellular networks) may not be applicable to multi-hop mobile environments since the data or request may need to go through multiple hops. As mobile clients in ad-hoc networks may have similar tasks and share common interest, cooperative caching, which allows the sharing and coordination of cached data among multiple clients can be used to reduce the bandwidth and power consumption.

To date there are some works in literature on cooperative caching in ad-hoc networks, such as consistency (Yin & Cao, 2004; Cao, Yin, & Das, 2004), placement (Zhang, Yin, & Cao, 2004; Nuggehalli, Srinivasan, & Chiasserini, 2003; Papadopouli & Schulzrinne, 2001), discovery (Takaaki & Aida, 2003), and proxy caching (Lau, Kumar, & Venkatesh, 2002; Friedman, Gradinariu, & Simon, 2004; Sailhan & Issarny, 2003; Lim, Lee, Cao, & Das, 2004, 2006). However, efficient cache replacement is not considered yet.

Cache management in mobile ad-hoc networks, in general, includes the following issues to be addressed:

1. The cache discovery algorithm that is used to efficiently discover, select, and deliver the requested data item(s) from neighboring nodes. In a cooperative architecture, the order of looking for an item follows local cache to neighboring nodes, and then to the original server.
2. The design of cache replacement algorithm—when the cache space is sufficient for storing one new item, the node places the item in the cache. Otherwise, the possibility of replacing other cached item(s) with the new item is considered.
3. Cache admission control—this is to decide what data items can be cached to improve the performance of the caching system.
4. The cache consistency algorithm, which ensures that updates are propagated to the copies elsewhere, and no stale data items are present.

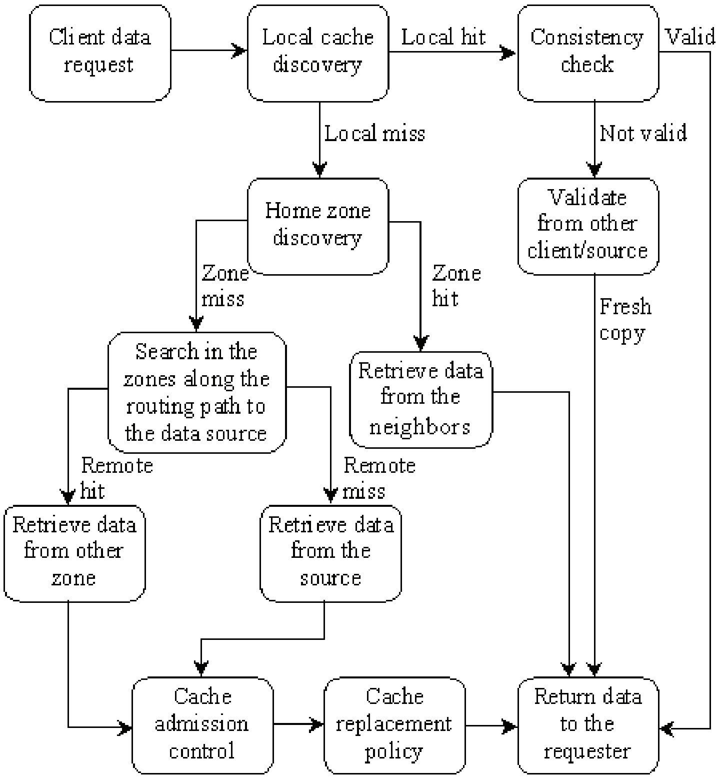
In this article, a *Zone Cooperative (ZC)* cache is proposed for mobile ad-hoc networks. Mobile nodes belonging to the neighborhood (zone) of a given node form a cooperative cache system for this node since the cost for communication with them is low both in terms of energy consumption and message exchanges. In ZC caching, each mobile node has a cache to store the frequently accessed data items. The cache at a node is a nonvolatile memory such as hard disk. The data items in the cache satisfy not only the node's own requests, but also the data requests passing through it from other nodes. For a data miss in the local cache, the node first searches the data in its cooperation zone before forwarding the request to the next node that lies on a path towards server. The caching scheme includes a discovery process and a cache management technique. The proposed cache discovery algorithm ensures that requested data are returned from the nearest node or server. As a part of cache management, cache admission control, Least Utility Value (LUV)-based replacement policy, and cache consistency technique are developed. The admission control prevents high data replication by enforcing a minimum distance between the same data item, while the replacement policy helps in improving the cache hit ratio and accessibility. Cache consistency ensures that clients only access valid states of the data.

SYSTEM ENVIRONMENT

The system environment is assumed to be an ad-hoc network where a mobile host accesses data items held as originals by other mobile hosts. A mobile host that holds the original value of a data item is called data server. A data source may be connected to the wired network. A data request initiated by a host is forwarded hop-by-hop along the routing path until it reaches the data source and then the data source sends back the requested data. Each mobile host maintains local cache in its hard disk. To reduce the bandwidth consumption and query latency, the number of hops between the data source/cache and the requester should be as small as possible. Most mobile hosts, however, do not have sufficient cache storage, and hence the caching strategy is to be devised efficiently. In this system environment, we also make the following assumptions:

- Assign a unique host identifier to each mobile host in the system. The system has a total of M hosts, and MH_i ($1 \leq i \leq M$) is a host identifier. Each host moves freely.
- We assign a unique data identifier to each data item located in the system. The set of all data items is denoted by $D = \{d_1, d_2, ..., d_N\}$, where N is the total number of data items and d_j ($1 \leq j \leq N$) is a data identifier. D_i denotes the actual data of the item with id d_i. Size of

Figure 1. Service of a client request in ZC caching strategy


Client data request
Local cache discovery
Local hit
Consistency check
Valid
Local miss
Not valid
Home zone discovery
Validate from other client/source
Zone miss
Zone hit
Fresh copy
Search in the zones along the routing path to the data source
Retrieve data from the neighbors
Remote hit
Remote miss
Retrieve data from other zone
Retrieve data from the source
Cache admission control
Cache replacement policy
Return data to the requester

data item d_i is s_i (in bytes)—that is, $s_i = |D_i|$. The origin of each data item is held by a particular data source.

- Each mobile host has a cache space of C bytes.
- Each data item is periodically updated at data source. After a data item is updated, its cached copy (maintained on one or more hosts) may become invalid.

ZONE COOPERATIVE CACHING

The design rationale of Zone Cooperative (ZC) caching is that it is considered advantageous for a client to share cache with its neighbors lying in the zone (i.e., clients that are accessible in one-hop). Mobile clients belonging to the cooperation zone of a given client then form a cooperative cache system for this client since the cost for communicating with them is low both in terms of energy consumption and message exchange. Figure 1 shows the behavior of ZC caching strategy for a client request.

When a data request is initiated at a client, it first looks for the item in its own cache. If there is a local cache miss, the client checks if the data item is cached in other clients within its home zone. When a client receives the request and has the data item in its local cache (i.e., a zone cache hit), it will send a reply to the requester to acknowledge that it has the data item. In case of a zone cache miss, the request is forwarded to the neighbor node along the routing path. Before forwarding a request, each client along the path searches the item in its local cache or zone as described above. If the data item is not found on the zones along the routing path (i.e., a remote cache miss), the request finally reaches the data source and the data source sends back the requested data.

When a client receives the requested data, a cache admission control is triggered to decide whether it should be brought into the cache. Inserting a data item into the cache might not always be favorable, because incorrect decision can lower the probability of cache hits. In ZC, the cache admission control allows a client to cache a data item based on the distance of data source or other client that has the requested data. If the original of the data resides in the same zone of the requesting client, then the item is not cached, because it is unnecessary to replicate a data item in the same zone since cached data can be used by closely located clients. In general, same data items are cached at least two hops away.

The ZC caching uses a simple weak consistency/invalidation model based on Time-To-Live (TTL), in which a client considers a cached copy up-to-date if its TTL has not expired.

The client removes the cached data when the TTL expires. A client refreshes a cached item and its TTL if a fresh copy of the item passes by.

UTILITY-BASED CACHE REPLACEMENT

The traditional cache replacement algorithms (e.g., LRU) might be unsuitable for ad-hoc networks due to non-uniform data size, varying distance between requester and data source, and frequent data updates.

We have developed a Least Utility Value-based cache replacement policy, where data items with the lowest utility are those that are removed from the cache. Four factors are considered while computing utility value of a data item at a client.

- **Popularity:** The access probability reflects the popularity of a data item for a host. An item with lower access probability should be chosen for replacement. At a client, the access probability A_i for data item d_i is given as

  $$A_i = a_i \Big/ \sum_{k=1}^{N} a_k ,$$

  where a_i is the mean access rate to data item d_i. a_i can be estimated by employing *sliding window* method of last K access times. We keep a sliding window of K most recent access timestamps $(t_i^1, t_i^2, ..., t_i^K)$ for data item d_i in the cache. The access rate is updated using the formula

  $$a_i = \frac{K}{t^c - t_i^K},$$

  where t^c is the current time and t_i^K is the timestamp of oldest access to item d_i in the sliding window. When fewer than K samples are available, all the samples are used to estimate a_i. To reduce the computational complexity, the access rates for all cached items are not updated during each replacement; rather the access rate for an item is updated only when the item is accessed. K can be as small as 2 or 3 to achieve the best performance. Thus the spatial overhead to store recent access timestamps is relatively small.
- **Distance:** Distance (δ) is measured as the number of hops between the requesting client and the responding client (source or cache). The greater the distance, the greater is the *utility* value of the data item. This is because caching data items that are further away saves bandwidth and reduces latency for subsequent requests.
- **Coherency:** A data item d_i is valid for a limited lifetime, which is known using the TTL_i field. An item that is valid for a shorter period should be preferred for replacement.
- **Size (s):** A data item with larger size should be chosen for replacement, because the cache can accommodate more items and satisfy more access requests.

Based on the above factors, the *utility$_i$* function for an item d_i is computed as:

$$utility_i = \frac{A_i.\delta_i.TTL_i}{s_i}.$$

The idea is to maximize the total utility value for the data items kept in the cache. For a cache of size C such that the size s_i of each data item d_i is very much less than C, the principle of optimality implies that the mobile client MH_x should always retain a set C_x of data items in its cache such that

$$\sum_{d_i \in C_x} utility_i$$

is maximized subject to

$$\sum_{d_i \in C_x} s_i \le C.$$

Maximizing the above objective function implies a minimization of the response time per reference. The task of LUV is to make this optimal decision for every replacement. A binary min-heap data structure is used to implement the LUV policy. The key field for the heap is the $utility_i$ value for each cached data item d_i. When the events of cache replacement occur, the root item of the heap is deleted. This operation is repeated until sufficient space is obtained for the incoming data item. Let N_c denote the number of cached items and S the victim set size. Every deletion operation has a complexity of $O(logN_c)$. An insertion operation also has an $O(logN_c)$ complexity. Thus the time complexity for every cache replacement operation is $O(SlogN_c)$.

CONCLUSION

Spurred by the progress of technologies and deployment at low cost, the use of ad-hoc networks is expected to be largely exploited for mobile computing, and no longer be restricted to specific applications (e.g., crisis applications as in military and emergency/rescue operations or disaster recovery). In particular, ad-hoc networks effectively support ubiquitous networking, providing users with network access in most situations. Data access in ad-hoc networks is an important issue where mobile nodes communicate with each other via short-range transmissions to share information. Caching of frequently accessed data in such an environment is a potential technique that can improve the data access performance and availability. In this article, we propose a novel scheme called *Zone Cooperative (ZC)* for caching in mobile ad-hoc networks. The scheme enables nodes to share their data which helps alleviate the longer average query latency and limited data accessibility problems in ad-hoc networks.

REFERENCES

Cao, G. (2002). On improving the performance of cache invalidation in mobile environments. *Mobile Networks and Applications, 7*(4), 291-303.

Cao, G. (2003). A scalable low-latency cache invalidation strategy for mobile environments. *IEEE Transactions on Knowledge and Data Engineering, 15*(5), 1251-1265.

Cao, G., Yin, L., & Das, C. (2004). Cooperative cache based data access framework for ad hoc networks. *IEEE Computer Magazine,* 32-39.

Chand, N., Joshi, R. C., & Misra, M. (2004). Broadcast based cache invalidation and prefetching in mobile environment. *Proceedings of HiPC* (LNCS 3296, pp. 410-419), Bangalore, India. Berlin: Springer-Verlag.

Chand, N., Joshi, R. C., & Misra, M. (2005). Energy efficient cache invalidation in a mobile environment. *International Journal of Digital Information Management (Special Issue on Distributed Data Management), 3*(2), 119-125.

Friedman, R., Gradinariu, M., & Simon, G. (2004). Locating cache proxies in MANETs. *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing* (pp. 175-186), Tokyo, Japan.

Lau, W. H. O., Kumar, M., & Venkatesh, S. (2002). Cooperative cache architecture in support of caching multimedia objects in MANETs. *ACM International Workshop on Wireless Mobile Multimedia* (pp. 56-63), Atlanta, GA.

Lim, S., Lee, W.-C., Cao, G., & Das, C. R. (2004). Performance comparison of cache invalidation strategies for Internet-based mobile ad hoc networks. *Proceedings of the IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* (pp. 104-113), FL.

Lim, S., Lee, W.-C., Cao, G., & Das, C. R. (2006). A novel caching scheme for improving Internet-based mobile ad hoc networks performance. *Ad Hoc Networks Journal, 4*(2), 225-239.

Nuggehalli, P., Srinivasan, V., & Chiasserini, C.-F. (2003). Energy-efficient caching strategies in ad hoc wireless networks. *Proceedings of the ACM Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)* (pp. 25-34), MD.

Papadopouli, M., & Schulzrinne, H. (2001). Effects of power conservation, wireless coverage and cooperation on data dissemination among mobile devices. *Proceedings of the ACM Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)* (pp. 117-127), CA.

Sailhan, F., & Issarny, V. (2003). Cooperative caching in ad hoc networks. *Proceedings of the International Conference on Mobile Data Management (MDM)* (pp. 13-28), Melbourne, Australia.

Takaaki, M., & Aida, H. (2003). Cache data access system in ad hoc networks. *Proceedings of the Vehicular Technology Conference (VTC)* (pp. 1228-1232), FL.

Yin, L., & Cao, G. (2004). Supporting cooperative caching in ad hoc networks. *Proceedings of IEEE INFOCOM* (pp. 2537-2547), Hong Kong.

Zhang, W., Yin, L., & Cao, G. (2004). Secure cooperative cache based data access in ad hoc networks. *Proceedings of the NSF International Workshop on Theoretical and Algorithmic Aspects of Wireless Ad Hoc, Sensor, and Peer-to-Peer Networks,* Chicago.

KEY TERMS

Access Point (AP): A transceiver in a wireless LAN that can connect a network to one or many wireless devices. APs can also bridge to one another.

Cache Consistency: A technique to ensure that the data at a client cache has same value as on the original server.

Cache Discovery: Searching the requested data in a MANET.

Cache Replacement: The process of eviction of an item from the cache when a new item is to be stored in the cache.

Caching: A technique where a copy of the remote data is stored locally to improve data availability and reduce access delay.

Cooperation Zone: One-hop neighbors of a mobile client form the cooperation zone for the client.

Data Server: A client that holds the original value of data.

Mobile Ad-Hoc Network: An autonomous network of mobile clients connected by wireless links where network topology may change rapidly and unpredictably.

Decision Analysis for Business to Adopt RFID

Koong Lin
Tainan National University of the Arts, Taiwan

Chad Lin
Edith Cowan University, Australia

Huei Leu
Industrial Technology Research Institute, Taiwan

INTRODUCTION

The spending for RFID (radio frequency identification) has been increasing rapidly in recent years. According to Gartner, global spending on RFID is likely to reach US$3 billion by 2010 (CNET, 2005). In addition, interests continue to grow for the adoption of this mobile computing and commerce device in many different types of applications (ABI, 2006). In 2005, Wal-Mart asked its top 100 suppliers to use RFID tags, and this had a profound effect on the projected growth of RFID technology as well as potential applications in the industrial, defense, and retails sectors (Albertsons, 2004).

However, very few studies have examined and evaluated the adoption of RFID options by the organizations. Organizations face various risks and uncertainties when assessing the adopted mobile technologies. Different organizations are likely to encounter different challenges and problems. This research aims to develop a mechanism that can help organizations to specify their risks and choose a suitable adoption alternative. This research has adopted the AHP (analytic hierarchy process) methodology to analyze the data, as it is useful for analyzing different RFID adoption alternatives and can assist organizations in predicting the possible issues and challenges when adopting RFID.

The objectives of this article are to: (1) describe basic components of a mobile computing and commerce device, RFID; and (2) explore the current practices, issues, and applications in this mobile technology.

BACKGROUND

RFID is a built-in wireless technology that incorporates a smart IC (integrated circuit) tag. It allows organizations to capture accurate information about the location and status of products, and track them as they move from the assembly line to the retail store (Albertsons, 2004). The three major components of RFID are: tags, readers, and software systems. RFID tags consist of silicon chips and antennas. Each tag uses an ID coding system and contains a unique serial number of a product. This enables the tag to store some information of the product. At present, the most well-known ID coding system is called EPC (electronic product code), which is formulated by MIT and used by Wal-Mart. The RFID EPC Network is constructed from the ONS (object name service), Savant (a middleware specific to RFID), and PML (Physical Markup Language) (AutoID, 2006; Lin et al., 2004).

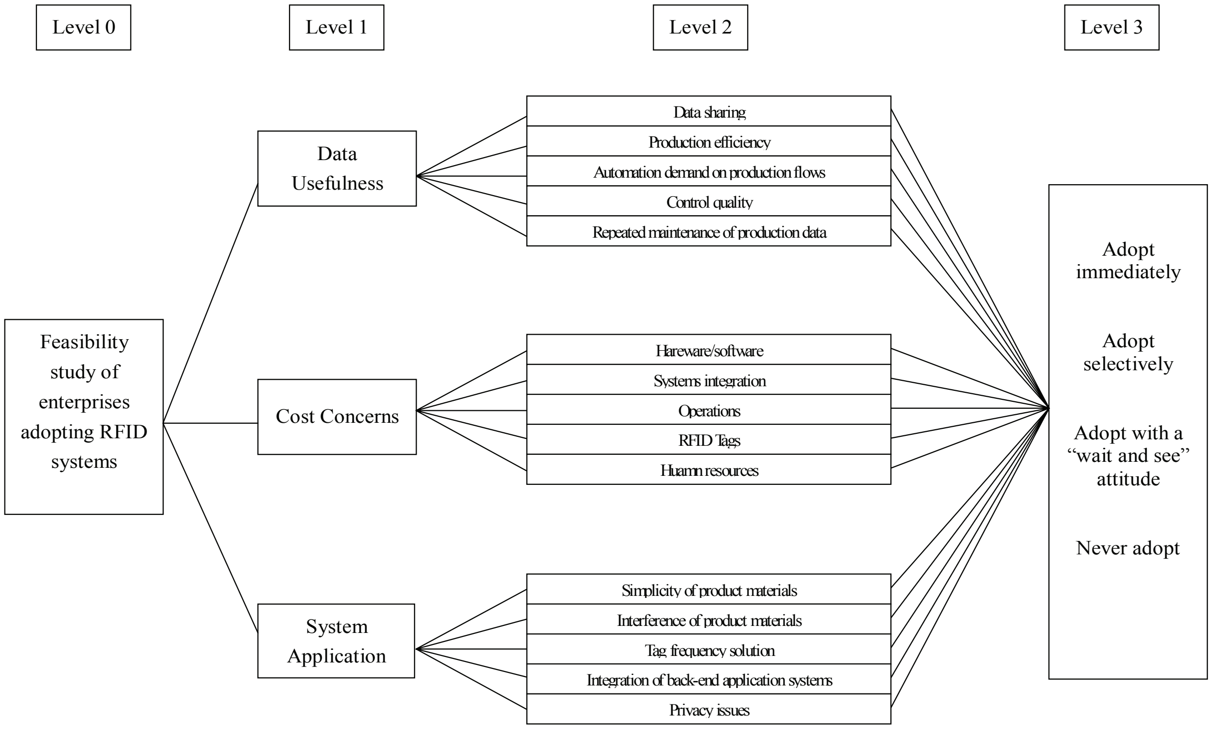
An RFID reader is used to communicate with RFID tags. In reading, the signal is sent out continually by the active tags. In interrogating, the reader sends a signal to the tags and listens. It can also send radio waves to energize the passive tags in order to receive their data. On the other hand, RFID software systems are the glue that integrates the RFID systems. The software systems manage the basic functions of the RFID reader and other components that route information to servers.

Organizations are using RFID in a number of data collection applications specific to their own industry, ranging from retail environments to hospitals, as well as what is currently being leveraged in warehouses to keep track of inventory and shipping. RFID has many advantages and can be deployed to assist organizations in improving global integration, as well as used as an effective tool in the areas of, for example, retail inventory tracking, customer relationship management, supply chain management, or any other situation where the tracking of the movement of goods or people is critical (Finkenzeller, 2003).

However, there are some business and technical problems and issues with the use of RFID technology (such as data sharing, data usefulness, accuracy, costs and benefits, security and privacy, and RFID standards) and this calls for further research. In essence, RFID requires collective and collaborative actions by stakeholders and organizations as a whole to ensure successful adoption and functioning of this technology. This is often affected by the divergent factors and perceptions of the internal and external stakeholders within an organization in the process of adopting RFIDs. For example, an organization must consider the potential costs in mastering collaborative planning and implementa-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure1. The research framework for evaluating RFID adoption


Level 0
Level 1
Level 2
Level 3
Feasibility study of enterprises adopting RFID systems
Data Usefulness
Cost Concerns
System Application
Data sharing
Production efficiency
Automation demand on production flows
Control quality
Repeated maintenance of production data
Hareware/software
Systems integration
Operations
RFID Tags
Huamn resources
Simplicity of product materials
Interference of product materials
Tag frequency solution
Integration of back-end application systems
Privacy issues
Adopt immediately
Adopt selectively
Adopt with a "wait and see" attitude
Never adopt

tion with its partners before attempting to share and use the RFID data (EPCglobal, 2006).

RESEARCH METHODOLOGY AND DESIGN

The AHP methodology is deployed to analyze the data collected and to build a decision support system. AHP was developed by Satty (1980) to reflect the way people actually think, and it continues to be the most highly regarded and widely used decision-making theory. In essence, AHP is a process that transforms a complicated problem into a hierarchical structure (Lin & Liu, 2005). By reducing complex decisions to a series of one-on-one comparisons and then synthesizing the results, AHP not only assists decision makers in arriving at the best decision, but also provides a clear rationale that it is the best (Lin et al., 2005).

The AHP methodology is useful for analyzing the RFID adoption alternatives as it can assist organizations in developing an integrated assessment of the entire organizational structure. AHP can also help to assess the inter-organizational issues among different divisions within an organization. Moreover, AHP can help to predict possible risks and challenges when adopting RFID so that the organizations are able to formulate appropriate strategies in order to minimize them (Satty, 1980; Wang & Yuan, 2001).

The following steps and considerations need to be taken into account when analyzing RFID adoption using AHP (Hair, Anderson, Tatham, & Black, 1998):

1. Issues may arise at the divisional level within an organization when adopting RFID.
2. The hierarchical structure for the organizations studied needs to be built using the collected data.
3. The questionnaire needs to be designed appropriately in order to identify and assess these issues.
4. Suggestions for improvement and/or alternatives need to be put forward in order to minimize the RFID adoption risks for the participating organizations.
5. Suggestions for improvement and/or alternatives also need to be communicated to all divisions within an organization, and alternatives need to be adjusted and revised accordingly.

Figure 1 depicts the AHP analysis hierarchical research framework for the evaluation of RFID adoption. Two types of questionnaires were designed for this research: (1) the expert questionnaire: to be completed by the RFID researchers and experts (expert evaluators); and (2) the industry questionnaires: to be completed by the RFID decision makers in those organizations that had a financial capital of at least US$1billion (industry evaluators).

- **Level 0:** The goals for this level were to conduct feasibility study of the industries involved with the adoption of RFID systems, as well as to identify and evaluate the importance of all major adoption factors (Lin et al., 2005).
- **Level 1:** After confirming the scope of the feasibility study for adoption of RFID in industries, three major factors were identified: data usefulness, cost concerns, and system application.
- **Level 2:** The three major factors identified in Level 1 were then decomposed into several criteria (in Level 2) which were evaluated according to their relative importance. These criteria were identified via interviews with the respondents, literature review, and surveys of industry characteristics.
- **Level 3:** Four alternatives were proposed for the adoption of RFID systems. The four proposed alternatives were: adopt immediately, adopt selectively, adopt with a "wait and see" attitude, and never adopt.

Data Analysis and Results

A total of 53 questionnaires were returned and the responses were analyzed using the AHP software. Significant differences were found between the responses from the RFID expert evaluators and industry evaluators. For example, "data sharing" was viewed by the expert evaluators as the most important issue for the "data usefulness" factor, whereas the "quality control" was the most important criterion for the industry evaluators.

Following this, the RFID experts were invited to provide their viewpoints on RFID techniques and theories. Using the AHP method, we found that both industry and expert evaluators had ranked the "cost concerns" of RFID as their number one factor for the adoption of RFID. The "application system" factor was ranked as the second most important concern by these evaluators, while they were least concerned about the "data usefulness" of RFID. In particular, costs of RFID tags and hardware were considered as the most important cost factor.

Weighting Analysis

The software, Expert Choice, was used to compute the weights from the responses. The consistency test was used to calculate the inconsistency ratio (IR) of the adoption criteria. All criteria have received consistent responses as their IR value is less than 0.1 (Satty, 1980). The AHP analysis indicated that both expert and industry evaluators ranked the adoption factors as follows: cost concerns (0.486 & 0.633), system applications (0.377 & 0.249), and data usefulness (0.137 & 0.118). The results are shown in Tables 1 and 2. Some research findings from Tables 1 and 2 are presented as follows.

Table 1. Adoption factor weightings of expert evaluators

Adoption Factors	Criteria	Weights
Data Usefulness (0.137)	Data sharing	0.034
	Production efficiency	0.023
	Automation demand on production flows	0.027
	Control quality	0.027
	Repeated maintenance of production data	0.025
Cost Concerns (0.486)	Hardware/software	0.126
	System integration	0.074
	Operations	0.085
	RFID tags	0.168
	Human resources	0.035
System Applications (0.377)	Simplicity of product materials	0.026
	Interference of product materials	0.064
	Tag frequency solution	0.037
	Integration of back-end application systems	0.082
	Privacy issues	0.167

Table 2. Adoption factor weightings of industry evaluators

Adoption Factors	Criteria	Weights
Data Usefulness (0.118)	Data sharing	0.017
	Production efficiency	0.029
	Automation demand on production flows	0.019
	Control quality	0.033
	Repeated maintenance of production data	0.020
Cost Concerns (0.633)	Hardware/software	0.155
	System integration	0.151
	Operations	0.116
	RFID tags	0.127
	Human resources	0.085
System Applications (0.249)	Simplicity of product materials	0.035
	Interference of product materials	0.032
	Tag frequency solution	0.030
	Integration of back-end application systems	0.058
	Privacy issues	0.095

Table 3. A comparison of weights by expert evaluators and industry evaluators

Criteria	Expert Weights	Industry Weights	Differences
RFID tag costs	0.168	0.127	0.041
Privacy issues	0.167	0.095	0.072
Hardware/software costs	0.126	0.155	0.029
Development/operation costs	0.085	0.116	0.031
Integration of back-end application systems	0.082	0.058	0.024
System integration costs	0.074	0.151	0.077
Interference of product materials	0.064	0.032	0.032
Tag frequency selection	0.037	0.03	0.007
Human resource costs	0.035	0.085	0.05
Data sharing	0.034	0.017	0.017
Automation of production flows	0.027	0.019	0.008
Quality control	0.027	0.033	0.006
Simplicity of product materials	0.026	0.035	0.009
Repeated maintenance of product data	0.025	0.02	0.005
Production efficiency	0.023	0.029	0.006

Data Usefulness

The responses from expert evaluators indicated that "data sharing" is far more important than the other adoption factors. "Data sharing" in RFID is generally defined as using a standardized data format to communicate between RFID supply chain suppliers. These respondents considered data format standardization as the most important issue in the process of adopting RFID. On the other hand, the responses from industry evaluators stated that "quality control" factor is their number one concern. This is not surprising given that businesses place great emphasis on business quality and service (RFID Journal, 2004).

Cost Concerns

As mentioned earlier, costs of RFID tags and hardware were considered as the most important cost factors from the expert evaluators' point of view, whereas the industry evaluators were more concerned about hardware and system integration costs. The high cost of RFID tags was viewed by the expert evaluators as the main obstacle for the adoption of RFID. On the other hand, industry evaluators were extremely concerned about the integration between the existing hardware and the new RFID systems. Most organizations would prefer to retain their current systems in order to avoid the system compatibility problem and increase the overall system performance (Traub, 2005).

Application System

Privacy was viewed as the most significant criterion for the "application system" factor by both the expert and industry evaluators. This was followed by the integration of a back-end system. This had reflected the fact that the privacy concern would likely affect the degree of trust during the future implementation of RFID. In addition, the problem with the integration of various back-end systems will also present security problems and other challenges for organizations to handle (Floerkemeier, 2003).

Table 3 depicts a comparison of weights by expert evaluators and industry evaluators.

In Table 4, criteria for RFID adoption were ranked according to their weights by both the expert and industry evaluators.

FUTURE TRENDS

Despite the fact that RFID has been widely adopted in many different fields at an increasing rate in recent years, the issues of security and privacy remain the key challenges

Table 4. Ranking of adoption criteria by expert evaluators and industry evaluators

Ranking	Expert		Industry	
	Criteria	Weights	Criteria	Weights
1	RFID tag costs	0.168	Hardware/software costs	0.155
2	Privacy issues	0.167	System integration costs	0.151
3	Hardware/software costs	0.126	RFID tag costs	0.127
4	Development/operation costs	0.085	Development/operation costs	0.116
5	Integration of back-end applications systems	0.082	Privacy issues	0.095
6	System integration costs	0.074	Human resource costs	0.085
7	Interference of product materials	0.064	Integration of back-end application systems	0.058
8	Tag frequency selection	0.037	Simplicity of product materials	0.035
9	Human resource costs	0.035	Quality control	0.033
10	Data sharing	0.034	Interference of product materials	0.032
11	Automation of production flows	0.027	Tag frequency selection	0.030
12	Quality control	0.027	Production efficiency	0.029
13	Simplicity of product materials	0.026	Repeated maintenance of product data	0.020
14	Repeated maintenance of product data	0.025	Automation of production flows	0.019
15	Production efficiency	0.023	Data sharing	0.017

in promoting the technology. Many industry experts have pointed out that the RFID-included objects should be targeted more efficiently by real-time tracking and instant management. However, the transmission of data is very vulnerable to eavesdropping because of the contact-less type of RFID remote retrieval. A primary security concern surrounding the RFID technology is the unsolicited tracking of consumer location and analyzing of their shopping habits or behavior. This is one important issue that needs to be addressed by RFID vendors.

In addition, the RFID technology can be further promoted by: (1) reducing the total costs of RFID; (2) resolving the interference problems; (3) improving the identification accuracy; (4) protecting the intellectual property rights; (5) establishing international standards for encoding systems, reader protocols, and programming environments; and (6) developing better software supports, including middleware, EPC information systems, and ONS design. Finally, it is expected that the RFID system will have a great impact on the way we work and live.

DISCUSSION AND CONCLUSION

The last section in the questionnaire asked the respondents to give scores to all four alternatives. The results have shown that both expert and industry evaluators preferred the second RFID adoption alternative (that is, to adopt RFID selectively).

The evaluators were also asked to weigh the relative importance of the three main factors as well as the criteria in each factor by a pair-wise comparison. The ranking of the adoption alternatives were obtained by multiplying the weighting given by expert evaluators and the weighting given by industry evaluators, by the scores obtained from the identified criteria in a matrix format. Some of the key findings from the AHP analysis are presented below:

1. When asked about the "data usefulness" factor, both expert and industry evaluators preferred the "adopt selectively" alternative. These evaluators indicated that the usefulness of data sharing and quick turnaround time in RFID could significantly improve the production efficiency and lower the costs of inventory if it was adopted in SCM (supply chain management) environments.
2. There were significant differences of opinion from both the expert and industry evaluators on the "cost concerns" factor. Expert evaluators preferred the "adopt selectively" alternative despite the high cost of implementing RFID. This was due to the fact that expert evaluators believed that the benefits of RFID would exceed the costs. On the other hand, industry experts were more conservative and preferred the

"adopt with a 'wait and see' attitude" due to the high costs of implementing RFID.

3. There were also significant differences of opinion from both the expert and industry evaluators on the "system application" factor. Expert evaluators were more concerned about the privacy issues than the industry evaluators.
4. Overall, the ranking for these four alternatives were: (1) adopt selectively, (2) adopt with a 'wait and see' attitude; (3) never adopt, and (4) adopt immediately. There was no difference in terms of the preference for these four alternatives by both the expert and industry evaluators.

In conclusion, organizations with the ability to quickly identify benefits, costs, and risks associated with RFID technologies and to effectively deploy them are more likely to gain competitive advantages from RFID. Regular evaluation of RFID technologies allows organizations to benefit from their implementation. Those organizations that regularly maintain strategic evaluation of RFID technologies can help to ensure that they will achieve RFID's true benefits.

REFERENCES

ABI. (2006). *RFID realism: Improving our ability to execute RFID projects with holistic solutions.* Retrieved from http://ww.abiresearch.com

Albertsons. (2004). The nation's second largest food and drug retailer, has launched its first RFID pilot project and announced that it will require its top 100 suppliers to tag pallets and cartons by April 2005. *RFID Journal.*

AutoID Labs. (2006). Retrieved from http://www.autoidlabs.org/

CNET. (2005). *Gartner sees RFID as $3 billion business by 2010.* Retrieved from http://www.news.com

EPCglobal. (2006). Retrieved from http://www.epcglobal.org.tw

Finkenzeller K. (2003). *RFID handbook: Fundamentals and applications in contactless smart cards and identification* (Trans. Rachel Waddington). Chichester, UK/New York: John Wiley & Sons.

Floerkemeier, C. (2003). *PML specification 1.0.* Retrieved from http//www.epcglobalinc.org/standards_technology/Secure/v1.0/PML_Core_Specification_v1.0.pdf

Hair, J. Jr., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate data analysis* (5th ed., p. 447). Upper Saddle River, NJ: Prentice Hall.

Lin, K. H., & Liu, L. (2005, December 23). On the development of digital TV commerce using the AHP mechanism. *Proceedings of the Technology & Business Forum 2005,* Hualien, Taiwan.

Lin, K. H., Chen, C. T., Ke, J. C., Leu, H., Yen, Y. C., & Yang, S. H. (2005, March 25-26). Using AHP technology to design a RFID adopting decision analysis system. *Proceedings of the 2005 Conference of Electronic Commerce and Digital Life (ECDL2005)*.

Lin, K. H., Chen, P., Juang, W., Dai, J., Jeng, W., Kuo, T., et al. (2004, June 10). Exploring the EPC network architecture for RFID technology application. *Proceedings of the 2004 Information Management Application and Development.*

RFID Journal. (2004). Target expects top vendor partners to apply tags to all pallets and cases and start shipping to select regional distribution facilities beginning late spring 2005. *RFID Journal,* (February 20).

Satty, T. L. (1980). *The Analystic Hierarchy Process.* New York: McGraw-Hill.

Traub, K. (2005). *ALE: A new standard for data access.* Retrieved from http://www.rfidjournal.com/article/articleview/1493/1/82/, 4/18

Wang, M., & Yuan, B. J. C. (2001). Application of Analytic Hierarchy Process (AHP) on the evaluation of alternative technologies—A case of digital TV receivers. *Tamsui Oxford Journal of Management Sciences, 17,* 29-42.

KEY TERMS

Analytic Hierarchy Process (AHP): Methodology developed by Satty (1980) to reflect the way people actually think; it continues to be the most highly regarded and widely used decision-making theory.

Electronic Product Code (EPC): It contains digits to identify the manufacturer, the product category, and the individual item.

Object Name Service (ONS): It looks up unique electronic product codes and points computers to information about the item associated with the code.

Physical Markup Language (PML): A method of describing products in a way computers can understand. PML, based on the widely accepted eXtensible Markup Language, is used to share information via the Internet in a format all computers can understand and use.

Radio Frequency Identification (RFID): A built-in wireless technology that incorporates a smart IC (integrated circuit) tag. The three major components of RFID are: tags, readers, and software systems.

RFID Reader: Used to communicate with RFID tags. In reading, the signal is sent out continually by the active tags. In interrogating, the reader sends a signal to the tags and listens.

RFID Software: RFID software systems are the glue that integrates the RFID systems. The software systems manage the basic functions of the RFID reader and other components that route information to servers.

RFID Tag: RFID tags consist of silicon chips and antennas. Each tag uses an ID coding system and contains a unique serial number of a product. This enables the tag to store some information of the product.

Definitions, Key Characteristics, and Generations of Mobile Games

Eui Jun Jeong
Michigan State University, USA

Dan J. Kim
University of Houston Clear Lake, USA

INTRODUCTION

In the emerging wireless environment of digital media communications represented as *ubiquitous* and *convergence,* rapid distribution of handheld mobile devices has brought the explosive growth of the mobile content market. Along with the development of the mobile content industry, mobile games supported by mobile features such as portability (mobility), accessibility (generality), and convenience (simplicity) have shown the highest growth rate in the world game market these days.

In-Stat/MDR (2004) and Ovum (2004) expect that the mobile games' annual growth rate between 2005 and 2009 will be around 50% in the United States and 30% in the world. According to KGDI (2005) and CESA (2005), compared to the rate of the whole game market (5%) of the world, it is about six times higher, and it exceeds the rate of video console (10%) and online games (25%). Mobile games thus are predicted to be one of the leading platforms in the world game market in 10 years' time. In addition, as the competition among game companies has been enhanced with the convergence of game platforms, mobile games are being regarded as a breakthrough for the presently stagnant game market, which has focused on heavy users.

However, due to the relative novelty of mobile games, there are a few visible barriers in the mobile game industry. First, definitions and terminologies and key characteristics related to mobile games are not clearly arranged as yet. Second, there is little research on the classification and development trends of mobile games. Therefore, this article is designed to contribute insights into these barriers in three ways. Firstly, the article provides narrow and broad definitions of mobile games. Secondly, key characteristics, platforms, and service types of mobile games are discussed. Finally, following the broad definition of mobile games, this article classifies mobile games as one to fourth generations and one pre-generation. Characteristics and examples of each generation are also presented.

DEFINITIONS OF MOBILE GAMES

Each country and each game research institution has different definitions and terminologies. The definition of mobile games is important because the functions of mobile devices are being converged with those of other devices. Mobile games—more precisely, mobile network games—are narrowly defined as *games conducted in handheld devices with network functionality.* The two key elements of this definition are *portability* and *networkability.* In this definition, mobile games are generally referred to as the games played in handheld mobile devices such as cell phones and PDAs with wireless communication functionality. In terms of portability and networkability, the characteristics of mobile games are different from other device platforms such as PC and console games, which do not have both portability and wireless capability. For example, Game Boy (GB) with no communication functionality was only regarded as a portable console device. However, this concept has lost some of its ground in the market since the advent of new mobile game devices from portable consoles such as Play Station Portable (PSP) and Nintendo Dual Screens (NDS), as wireless networked games began to be serviced through the new mobile game devices.

Mobile games can be broadly defined as *embedded, downloaded, or networked games conducted in handheld devices such as mobile phones, portable consoles, and PDAs.* The key element of this concept is portability: all games in portable devices can be thought of as mobile games without regard to wireless functions. Therefore, this concept expands mobile games by including video games in portable consoles and embedded games in general portable devices such as PDAs, calculators, and dictionaries. As most game devices have been adopted with wireless networking functions, this definition becomes more powerful in game markets.

Recently, the narrow definition of mobile games has been generally used. However, since the meaning of *mobile* includes that of *portable and network (either wired or wireless function is embedded),* the broad definition of mobile

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

games including portable game-dedicated devices such as GBs and PSPs should be used. This definition is more persuasive in the present and future game market. For instance, the competition between Nokia's N-gage (i.e., a cell phone integrating the functions of MP3 and games) and Sony's PSP (i.e., a portable game machine including functions of MP3 and networking) is for the preoccupation of a future mobile platform.

KEY CHARACTERISTICS, PLATFORMS, AND SERVICE TYPES

Characteristics and Limitations of Mobile Games

Mobile games are differentiated from other platform games such as console, PC, and arcade games in terms of their portability, accessibility, networkability, and simplicity. Owing to the *portability* (i.e., mobility), users can play games anytime. This characteristic has attracted many light users, who play simple games such as puzzle, card, or word games, because these games can be played in one's spare time in a short amount of time. Compared to players in other genres such as role playing games (RPGs) and simulation games that require a long time to play, light users vary broadly in terms of age, and many women players also belong to this group. This is one of the strongest potentials of mobile games. The second characteristic of mobile games is *accessibility.* This can be defined as to the extent one can use a mobile device to play games at anytime and at anyplace. Console games are restricted to owners who have console machines and who want to enjoy games for a long time in a particular place. Likewise, most PC games and arcade games need to be somewhere in front of game devices with network facilities. However, mobile games—especially using mobile devices—are easy to access, because people almost always bring those devices anywhere and can download games anywhere as long as wireless networks are available. The third characteristic is *networkability.* Through wired or wireless connections, online games and console games are transplanted into mobile games to facilitate game usages. For example, some online games are linked to mobile games, so those games can be used both in PCs and mobile devices: game users can play the games with no limits in terms of location, machine, and time. Furthermore, mobile game users can play multi-user real-time games such as MMORPG (massively multiplayer online role-playing game) and real-time strategy (RTS) games. The final characteristic is their *simplicity* to use: mobile devices are simpler to handle than other platform machines. In addition, it is much easier to acquire the skills of the games and use them than those of other platforms.

Because of these characteristics, mobile games develop faster than other platform games. According to W2F (2003) and KGDI (2005), the development of a PC or console game usually takes at least two years to develop with more than 20 trained people and about $3 million. But in mobile games, about three to six months are spent with five people and less than $150,000. This is why the initial market entry barrier of mobile games is lower than that of other platform game markets. However, the average lifecycle of mobile games is less than six months, and value chains are more complex than those of other platform games.

Despite the major advantages of mobile games, there are drawbacks in some points. The most essential point is from not-unified platforms. With Internet and console games, converting of original games is not necessary, because the original games can be available in any PC via the Internet. However, mobile games should be converted to make them fit to other platforms, even in the same area. In other words, the conversion is necessary for service to be available in other mobile devices. The second is small screens and low capable devices. Although 3D networked games are being serviced, small screens and monotonous sounds are not sufficient to maximize the feelings of presence for users, and mobile game devices still do not have enough capacity to download high-capacity games through mobile networks.

Mobile Game Platforms

Mobile platforms function as game engines by running applications: a game engine is the core code handling the basic functionality of a game. Each mobile device has its own platform, so developers make games based on the formats of those platforms. With the development of platforms, downloaded, 3D games, and more advanced games are now serviced. These platforms are either freely opened or purchased with license fees. Platform holders have tried to expand their platforms, because the prevalence of their platforms implies a strong influence in mobile markets. These days, Java is the most influential platform both in mobile phone games and in handset manufacture. The Java 2 Micro Edition (J2ME) is a freeware version of Java; Execution Engine (ExEn) and Mophun are also freeware platforms distributed mainly in Europe. Brew is the licensed platform mainly used in the United States, Japan, and Korea. Different from mobile phone games, portable console games such as GB, N-Gage, PSP, and NDS have their own development tools for the platforms. Developers who want to make mobile games in portable consoles should use such development kits with the charge of license fees. Since developers adopt more prevailing game kits for the better benefit of their games, the market prevalence of console platforms is parallel with the amount of license fees for portable console manufacturers.

Mobile Game Service Types

With the development of mobile service technologies, mobile game services have evolved from single/embedded to multiplayer networked games. Single/embedded are games with which just one player can play without network services. These embedded games are still used in many mobile devices as a service for device customers. Message-based are games using short message service (SMS). These types are played in wireless network environments through WAP (Wireless Application Protocol) browsing environments, but these games are shifting into multi-media message service (MMS) with high capabilities providing enhanced messaging services such as graphics, sound, animation, cartoons, and texts. Downloaded games have been developed with the advent of download platforms such as Java, ExEn, and Mophun. These games have been taken usually from mobile portals managed by mobile network operators, with charges based on both content and traffic fees. Networked games are the newest type of mobile games which are activated with the advent of the flat sum systems.

GENERATIONS OF MOBILE GAMES

From the broad definition of mobile games perspective, portable console machines were the first mobile devices that emerged in the 1970s. These games have been categorized as console games because most hit games were published by console game companies such as Nintendo, Sega, and Sony, and game users were the same as those of console games. However, they have expanded their ranges into color graphic games in the 1990s and mobile network games in the 2000s, so users of such games are no longer limited to young boys not yet in their teens. Following the broad definition of mobile games, this article includes portable console games as a part of mobile games. Portable console games, made before the advent of network portable console games around 2003, are regarded as "portable embedded games," which are categorized as the pre-generation of mobile games.

The first wireless mobile phone game, *Snake,* was serviced as a text-based (or early-graphic) game in 1997. However, today's state-of-the-art games are 3D, fully networked multi-user games with high definitions in wide color screens. As the development of mobile interfaces and network functionalities continues, mobile games can be divided into four generations and one pre-generation. These generations are categorized by stand-alone (off-line) or networked, text-based or graphic-based, and 2D or 3D graphics.

Pre-generation (Pre-G) refers to portable console games that are played in standalone portable devices. In the 1970s, these games were all embedded in only-one-game-use portable game machines such as *Auto Race, Merlin,* and *Missile Attack* by American vendors such as Mattel, Entex, and Tomy. However, in the 1980s, both embedded and cartridge usable games were pervaded with the initiatives of Japanese game companies such as Nintendo and Bandai. From 1989, portable console games were converged into the Nintendo Game Boy era with cartridge games. In the mid-1990s, these games were serviced with color graphic games. With various games usually transplanted from console games, these portable console games flourished with the development of console games.

The *first generation (1G)* refers to text-based mobile phone games like puzzle games. They had been usually serviced by wireless application protocol (WAP) from 1998, and most of them are single-player embedded games. Some early-graphic games were embodied by white and black dots. These games spread until around 2001 when mobile platforms such as Java, Brew, and ExEn began to spread for the development of mobile graphic games. The *second generation (2G)* refers to graphic games. Developers transported popular games in PC or console games into mobile devices. At first, all graphic services were 2D white and black, but from around 2002, color phones rapidly spread with color graphic games, and some functions such as chatting and reviewing were added. With the prevalence of download platforms, downloaded games generally began to be provided by mobile portals. Traditional board games such as card and chess games were also translated into mobile graphic games in this generation with the concept of licensed games.

The *third generation (3G)* refers to networked games with simple network functionalities. Around 2003, most games were 2D graphic: 3D games were just a state of experiment. Network functionality was not fully serviced, because of the high cost of network use and low capabilities of mobile devices. However, owing to network capability, new games such as various simulations, multi-user role-playing games (MRPG), and location-based service (LBS) games were firstly developed in this period. Additionally, new mobile devices such as N-gage, PSP, and NDS had changed the traditional concepts of mobile games with the mixture of wireless and networked game services. These devices are estimated to have promoted the degree of mobile games as much as that of console games. With the prevalence of device convergences between mobile and console devices, from this generation there is no accurate difference among game genres. The *fourth generation (4G)* games refer to 3D and full networked games such as massively multi-user online role playing games (MMORPG). In addition, around 2004, full 3D mobile game services began to be serviced, and many developers joined the development of 3D network games. With the spread of new 3D graphic mobile phones and flat sum systems, 3D networked games are steadily gaining their shares in game markets. Table 1 illustrates the mobile game generations, key characteristics of each generation, and examples.

Table 1. Generations of mobile games

	Outset	Characteristic Game Genres	Examples
Pre-G	1970s	Portable console games Portable console color games (Embedded or cartridge games)	*Auto Race, Football* *Bomberman* *Pac Man, Tetris (Console)*
1G	1997/1998	Text-based games (early-graphic) WAP games	*Snake* *Dataclash, Gladiator*
2G	2001/2002	Downloaded games (Java, Brew) 2D color graphic games	*Tetris, Chess Mobile* *Samurai Romanesque*
3G	2003/2004	Portable console network games Half network games 3D graphic games	*Pokemon Ruby* *Badlands, Samgukji* *3D Pool*
4G	2004/2005	Full 3D graphic games Full 3D network games	*3D Golf, 3D Bass Fishing* *Homerun King Mobile*

CONCLUSION

With the expansion of a convergence and ubiquitous environment, the range of mobile games has grown to include all games available in handheld devices with portability. At first, mobile games were regarded as embedded single-user games. However, through the development of network and graphic technology, mobile games have been played as both full network games with multi-users and 3D graphic games with high-definition devices. So, many high-capability games such as MMORPG and multi-user simulation games have been adopted in mobile devices with high-speed network capability. In addition, with the convergence of game devices, the boundary between mobile phones and console devices has been eroded, while games in PC and console machines have been transformed into mobile games. Due to the accessibility, portability, and ease of use, mobile games have a wide range of users and do not impose limitations in age, sex, and social status. Traditionally game users were usually young males, but when it comes to mobile games, the game users are more diversified: not only young boys, but also elderly people and women are joining in on the new mobile gaming era. With the development of mobile technology, diversification of mobile content services, and generalization of mobile game users, mobile games will continue to gain more power within game markets.

Mobile games are summarized along with taxonomies. In addition, recent trends with game application areas will be discussed in the next article, "Mobile Games Part II: Taxonomies, Applications, and Trends."

REFERENCES

CESA. (2005). *2005 CESA game white paper.* Tokyo: Computer Entertainment Suppliers' Association.

DeMaria, R., & Wilson, J. L. (2002). *High score: The illustrated history of electronic games.* Berkeley: Osborne Media Group.

Entertainment Software Association. (2002). *Top ten industry facts.* Retrieved from www.theesa.com/pressro om.html

Ermi, L., & Mäyrä, F. (2005). Challenges for pervasive mobile game design: Examining players' emotional responses. *Proceedings of the ACM SIGCHI International Conference on Advances in Computer Entertainment Technology* (pp. 47-55), Valencia, Spain.

Hall, J. (2005). Future of games: Mobile gaming. In J. Raessens & J. Goldstein (Eds.) *Handbook of computer game studies* (pp. 47-55). Cambridge, MA: MIT Press.

In-Stat/MDR. (2004). Mobile gaming services in the U.S., 2004-2009. *In-Stat/MDR,* (August).

KGDI. (2005). *2005 game white paper.* Seoul: Korea Game Development & Promotion Institute.

Newman, J. (2004). *Videogames.* London: Routledge.

Nokia. (2003). *Introduction to mobile game development, Nokia Corporation.* Retrieved from www.forum.nokia.com/html_reader/main/1,,2768,00.html

Ovum. (2004, December). *Ovum forecasts global wireless market.* Ovum.

Ring, L. (2004). *The mobile connection: The cell phone's impact on society.* San Francisco: Morgan Kaufmann.

Taylor Nelson Sofres. (2002). *Wireless and Internet technology adoption by consumers around the world.* Retrieved from www.tnsofres.com/IndustryExpertise/IT/WirelessandInternetAdoption byConsumersAroudtheWorldA4.pdf

W2F. (2003, October). *Winning and losing in mobile games.* W2F Limited.

KEY TERMS

Device Platform: A device such as a cell phone, PDA, PC, or console machine through which games can be played.

Local-Based Service (LBS) Game: A mobile network game played within a local place around a telecommunication base with the information of a user's position.

Massively Multi-player Online Game (MMOG): A game where a huge number of users can play simultaneously based on their roles or missions.

Mobile Game Platform: Core code handling of the basic functionality of a game such as downloading, networking, or activating 3D graphics.

Mobile Game: An embedded, downloaded, or networked game conducted in a handheld device such as a mobile phone, portable console, or PDA.

Portable Console (Device): Handheld console machine such as PSP, NDS, and GBA with portable capabilities.

Role Playing Game (RPG): A game where a gamer takes a role and uses items in accomplishing missions or quests.

3D Network Game: A game played in connection with other users, using 3D graphics.

Design Methodology for Mobile Information Systems

Zakaria Maamar
Zayed University, UAE

Qusay H. Mahmoud
University of Guelph, Canada

INTRODUCTION

Mobile information systems (MISs) are having a major impact on businesses and individuals. No longer confined to the office or home, people can use devices that they carry with them, along with wireless communication networks, to access the systems and data that they need. In many cases MISs do not just replace traditional wired information systems or even provide similar functionality. Instead, they are planned, designed, and implemented with the unique characteristics of wireless communication and mobile client use in mind. These unique characteristics feature the need for specific design and development methodologies for MISs. Design methods allow considering systems independently of the existing information technologies, and thus enable the development of lasting solutions. Among the characteristics that a MIS design method needs to consider, we cite: unrestricted mobility of persons, scarcity of mobile devices' power-source, and frequent disconnections of these devices.

The field of MISs is the result of the convergence of high-speed wireless networks and personal mobile devices. The aim of MISs is to provide the ability to compute, communicate, and collaborate anywhere, anytime. Wireless technologies for communication are the link between mobile clients and other system components. Mobile client devices include various types, for example, mobile phones, personal digital assistants, and laptops. Samples of MIS applications are mobile commerce (Andreou et al., 2002), inventory systems in which stock clerks use special-purpose mobile devices to check inventory, police systems that allow officers to access criminal databases from laptops in their patrol cars, and tracking information systems with which truck drivers can check information on their loads, destinations, and revenues using mobile phones. MISs can be used in different domains and target different categories of people.

In this article, we report on the rationale of having a method for designing and developing mobile information systems. This method includes a conceptual model, a set of requirements, and different steps for developing the system. The development of a method for MISs is an appropriate response to the need of professionals in the field of MISs. Indeed, this need is motivated by the increased demand that is emerging from multiple bodies: wireless service providers, wireless equipment manufacturers, companies developing applications over wireless systems, and businesses for which MISs are offered. Besides all these bodies, high-speed wireless data services are emerging (e.g., GPRS, UMTS), requiring some sort of new expertise. A design and development method for MISs should support professionals in their work.

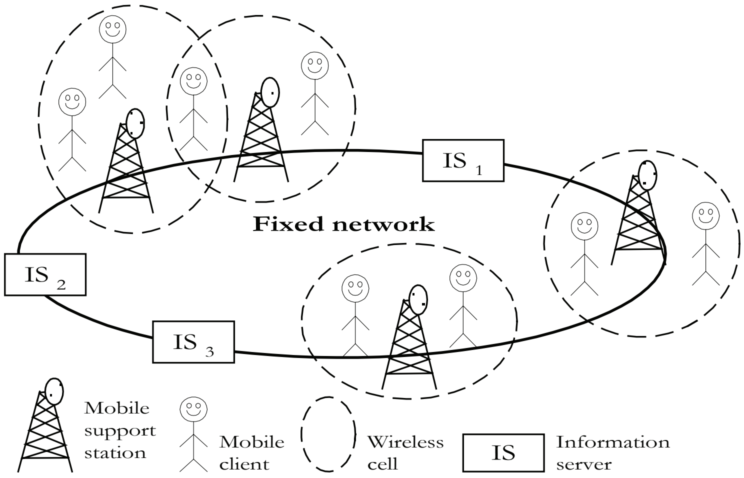
MOBILE COMPUTING MODEL

The general mobile computing model in a wireless environment consists of two distinct sets of entities (Figure 1): mobile clients (MCs) and fixed hosts. Some of the fixed hosts, called mobile support stations (MSSs), are enhanced with wireless interfaces. An MSS can communicate with the MCs within its radio coverage area called wireless cell. An MC can communicate with a fixed host/server via an MSS over a wireless channel. The wireless channel is logically separated into two sub-channels: an uplink channel and a downlink channel. The uplink channel is used by MCs to submit queries to the server via an MSS, whereas the downlink channel is used by MSSs to disseminate information or to forward the responses from the server to a target client. Each cell has an identifier (CID) for identification purposes. A CID is periodically broadcasted to all the MCs residing in a corresponding cell.

The wireless application protocol (WAP) is a technology that plays a major role in the field deployment of the mobile computing model (Open Mobile Alliance). WAP is an open, global specification that empowers users with mobile devices to easily access and interact with information and services instantly. It describes how to send requests and responses over a wireless connection, using the wireless session protocol (WSP), which is an extended and byte-coded version of HTPP 1.1. A WSP request is sent from a mobile device to a WAP gateway/proxy to establish an HTTP session with the target Web server. Over this session, the WSP request, converted into HTTP, is sent. The content, typically presented

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Representation of the mobile computing model



in the Wireless Markup Language (WML), is sent back to the WAP gateway, where it is byte-coded and sent to the device over the WSP session.

REQUIREMENTS FOR MISs

The role of an MIS is to provide information to mobile users through wireless communication networks. Two aspects are highlighted here: information and network. Information has to be available, taking into account terrain and propagation techniques. Plus, the information exchange has to be secured. A security problem inherent to all wireless communication networks consists of third parties being able to easily capture the radio signals while in the air. Thus, appropriate data protection and privacy safeguards must be ensured. Regarding the network element, this latter needs to consider failure cases and recover from them.

1. **Information Availability Requirement:** This illustrates the need for a user to have uninterrupted and secure access to information on the network. Aspects to consider are: survivability and fault tolerance, ability to recover from security breaches and failures, network design for fault tolerance, and design of protocols for automatic reconfiguration of information flow after failure or security breach.
2. **Network Survivability Requirement:** This illustrates the need to maintain the communication network "alive" despite of potential failures. Aspects to consider are: understand system functionality in the case of failures, minimize the impact of failures on users, and provide means to overcome failures.
3. **Information Security Requirement:** This illustrates the importance of providing reliable and unaltered information. Aspects to consider are: confidentiality to protect information from unauthorized disclosure, and integrity to protect information from unauthorized modification and ensure that information is accurate, complete, and can be relied upon.
4. **Network Security Requirement:** This illustrates the information security using network security. Aspects to consider are: confidentiality, sender authentication, access control, and identification.
5. **Additional Requirements of MIS Have Been Put Forward:** Indeed, the increasing reliance and growth in information-based wireless services impose three requirements—availability, scalability, and cost efficiency—on the services to be provided. Availability means that users can count on accessing any wireless service from anywhere, anytime, regardless of the site, network load, or device type. Availability also means that the site provides services meeting some measures of quality such as short, acceptable, and predictable response time. Scalability means that service providers should be able to serve a fast-growing number of customers with minimal performance degradation. Finally, cost effectiveness means that the quality of wireless services (e.g., availability, response time) should come with adequate expenditures in IT infrastructure and personnel.

CHALLENGES AND POSSIBLE SOLUTIONS IN MISs

The requirements discussed above pose several crucial challenges, which must be faced in order for MIS applications to function correctly in the target environment.

- **Transmission Errors:** Messages sent over wireless links are exposed to interference (and varying delays) that can alter the content received by the user, the target device, or the serer. Applications must be prepared to handle these problems. Transmission errors may occur at any point in a wireless transaction and at any point during the sending or receiving of a message. They can occur after a request has been initiated, in the middle of the transaction, or after a reply has been sent.
- **Message Latency:** Message latency, or the time it takes to deliver a message, is primarily affected by the nature of each system that handles the message, and by the processing time needed and delays that may occur at each node from origin to destination. Message latency should be handled, and users of wireless applications should be kept informed of processing delays. It is especially important to remember that a message may be delivered to a user long after the time it is sent. A long delay might be due to coverage problems or transmission errors, or the user's device might be switched off or have a dead battery.
- **Security:** Any information transmitted over wireless links is subject to interception. Some of that information could be sensitive, like credit card numbers and other personal information. The solution needed really depends on the level of sensitivity.

Here are some practical hints useful to consider when developing mobile applications. These hints back the development of the proposed method for designing mobile information systems.

- **Understand the Environment and Do Some Research Up Front:** As with developing any other software application, we must understand the needs of the potential users and the requirements imposed by all networks and systems the service will rely on.
- **Choose an Appropriate Architecture:** The architecture of the mobile application is very important. No optimization techniques will make up for an ill-considered architecture. The two most important design goals should be to minimize the amount of data transmitted over the wireless link, and to anticipate errors and handle them intelligently.
- **Partition the Application:** Think carefully when deciding which operations should be performed on the server and which on the handheld device. Downloadable wireless applications allow locating much of an application's functionality of the device; it can retrieve data from the server efficiently, then perform calculations and display information locally. This approach can dramatically reduce costly interaction over the wireless link, but it is feasible only if the device can handle the processing the application needs to perform.
- **Use Compact Data Representation:** Data can be represented in many forms, some more compact than others. Consider the available representations and select the one that requires fewer bits to be transmitted. For example, numbers will usually be much more compact if transmitted in binary rather than string forms.
- **Manage Message Latency:** In some applications, it may be possible to do other work while a message is being processed. If the delay is appreciable—and especially if the information is likely to go stale—it is important to keep the user informed of progress. Design the user interface of your applications to handle message latency appropriately.
- **Simplify the Interface:** Keep the application's interface simple enough that the user seldom needs to refer to a user manual to perform a task. To do so, reduce the amount of information displayed on the device, and make input sequences concise so the user can accomplish tasks with the minimum number of button clicks.

PROPOSED METHOD

The first step towards a successful wireless implementation project is a thorough business analysis, which serves as the backbone of any project bearing a fruitful return of investment. The analysis ensures that the project's requirements result in a wireless implementation that will successfully meet users' expectations and needs. Next, determinations about development features, approach, and constraints are made. This ensures that the wireless implementation is a good fit with the planned usage and infrastructure of the company. Finally, a choice needs to be made with regard to the software and hardware systems.

Business Analysis

When considering a wireless implementation, several questions have to be considered:

- What are the overall goals for implementing wireless services?
- What are the new markets to be targeted?
- What are the goals for giving mobile customer/staff wireless remote access?
- What technologies are currently in place towards supporting a wireless enterprise?
- Is interactivity important to the company?
- What current functions are suitable for wireless use?

- How prepared is the infrastructure to develop and host wireless applications?
- Are resources available to develop, implement, and support the wireless project?
- Is it more economical to have a wireless solution compared to a wired one?

Development

There are several factors to take into account when determining the wireless development solution. Indeed, MISs are expanding rapidly and changing from largely voice-oriented to increasingly data and multimedia systems.

- **Online vs. Off-Line:** On one hand, online applications include functions that require continuous connectivity, for example, looking for an inventory status or checking for available flights. Online wireless applications require real-time connectivity to be effective and useful. On the other hand, offline applications do not require real-time connectivity. Instead, they reside locally on a particular wireless device and are always available for use, but not always in real time. In addition, their use is limited to that particular device.
- **Screen Size:** With a much smaller display area than traditional desktops/laptops, it is important to fit in as much user-required functionality as possible, while trying to format the information in such a way that it appears attractive and appealing to end users.
- **Color:** Not all wireless devices support color, and some of them support a broader palette than others. Therefore, it is important to consider each device's color support, especially if multicolor content is to be provided, such as maps or advertisement banners.
- **Ergonomics:** Wireless devices vary widely in their standards and capabilities from one to the other. Which devices should be supported, which ones are best suited for the application's needs, and can all these devices be supported at the same time?

The above-listed questions have to be associated with a development lifecycle of the MIS. We advocate the consideration of four stages to constitute that lifecycle:

1. Requirements Stage
    - Identify key information that users need when mobile.
    - Establish use-case scenarios for such information.
    - Illustrate these scenarios to users for validation purposes.
2. Analysis Stage
    - Analyze and compare similar systems (wired or mobile) to the future mobile system.
    - Identify the elements that are directly linked to wireless aspects.
    - Highlight features of different wireless devices that the future wireless system will support.
    - Identify the needed wireless communication technologies as well as the network topologies on which the future wireless system will be built.
    - Analyze the various technologies to get users' queries and return responses (e-mails, SMS, WML, etc.). Dempsey and Donnelly (2002) listed some of the key features of an m-interface: usability, intelligent and personalized services, security, consultation capabilities, and pervasive and flexible payment mechanisms.
    - Analyze security and scalability problems.
3. Design Stage
    - Analyze security and scalability problems.
    - Use existing information resources or tailor them for mobile use.
    - Develop the architecture of the future application at data and process levels.
    - Discuss the location of data and processes, and who is in charge of maintaining these data and implementing these processes.
    - Provide solutions to potential security and scalability problems.
4. Implementation Stage
    - Develop and test the new application using for example Java 2 Micro Edition (J2ME) platform (http://java.sun.com).
    - Deploy the new application on the field.

CONCLUSION

In this article, we overviewed our vision of the importance of having a dedicated design method for mobile information systems. This importance is motivated by the continuous pressure on the professionals of MISs, who are demanded to put new solutions according to the latest advances in the mobile field. For instance, it is no longer accepted to postpone operations just because there is no connection to a fixed computing desktop. Mobile devices are permitting new opportunities when it comes to banking, messaging, and shopping, just to cite a few.

REFERENCES

Andreou, A. S., Chrysostomou, C., Leonidou, C., Mavromoustakos, S., Pitsillides, A., Samaras, G., Samaras, C., & Schizas, C. (2002). Mobile commerce applications and services: A design and development approach. *Proceedings* *of the 1st International Conference on Mobile Business (MBusiness 2002)*, Athens, Greece.

Bellavista, P., Corradi, A., & Stefanelli, C. (2002). The ubiquitous provisioning of Internet services to portable devices. *IEEE Pervasive Computing, 1*(3).

Campo, C. (2002). Service discovery in pervasive multi-agent systems. *Proceedings of the 1st International Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices* (in conjunction with *AAMAS'2002*), Bologna, Italy.

Castano, A., Ferrara, S., Montanelli, S., Pagani, E., & Rossi, G. P. (2003). Ontology-addressable contents in P2P networks. *Proceedings of the 1st Workshop on Semantics in Peer-to-Peer and Grid Computing* (in conjunction with *WWW'2003*), Budapest, Hungary.

Dempsey, S., & Donnelly, W. (2002). Identifying the building blocks of mobile commerce. *Proceedings of the 1st International Conference on Mobile Business (MBusiness'2002)*, Athens, Greece.

Elsen, I., Hartung, F., Horn, U., Kampmann, M., & Peters, L. (2001). Streaming technology in 3G mobile communication systems. *IEEE Computer, 34*(9).

Jose, R., Moreira, A., & Rodrigues, H. (2003). The AROUND architecture for dynamic location-based services. *Mobile Networks and Applications, 8*(4).

Karakasidis, A., & Pitoura, E. (2002). DBGlobe: A data-centric approach to global computing. *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops (ICDCSW 2002)*, Vienna, Austria.

Konig-Ries, B., & Klein, M. (2002). Information services to support e-learning in ad-hoc networks. *Proceedings of the 1st International Workshop on Wireless Information Systems* (in conjunction with *ICEIS 2002*), Ciudad Real, Spain.

Maamar, Z., Ben-Younes, K., & Al-Khatib, G. (2003). Scenarios of supporting mobile users in wireless networks. *Proceedings of the 2nd International Workshop on Wireless Information Systems* (in conjunction with *ICEIS 2003*), Angers, France.

Open Mobile Alliance. (2005). Retrieved June 2005 from http://www.wapforum.org

Raghu, T. S., Ramesh, R., & Whinston, A. B. (2002). Next steps for mobile entertainment portals. *IEEE Computer, 35*(5).

Ratsimor, O., Chakraborty, D., Tolia, S., Kushraj, D., Kunjithapatham, A., Gupta, G., Joshi, A., & Finin, T. (2002). Allia: Alliance-based service discovery for ad-hoc environments. *Proceedings of the 2nd ACM Mobile Commerce Workshop* (in conjunction with *MOBICOM 2002*), Atlanta, GA.

KEY TERMS

Cell: A geographic area defining the range in which a mobile support station supports a mobile client. Each cell has a cell identifier (CID) that uniquely describes it.

General Packet Radio Service (GPRS): A mobile data service available to users of global system for mobile communications (GSM) users.

Java 2 Micro Edition (J2ME): An edition of the Java platform for developing applications that can run on consumer wireless devices such as mobile phones.

Mobile Client (MC): A user with a handheld wireless device that is able to move while maintaining its connection to the network.

Mobile Information System (MIS): A computing information system designed to support users of handheld wireless devices.

Mobile Support Station (MSS): A static host that facilitates the communication with mobile clients. An MSS supports mobile clients within a geographic area known as a cell.

Short Message Service (SMS): A service for sending text messages, up to 160 characters each, to mobile phones.

Universal Mobile Telecommunications System (UMTS): A third-generation (3G) mobile phone technology.

Wireless Application Protocol (WAP): A standard specification for enabling mobile users to access information through their handheld wireless devices.

Wireless Markup Language (WML): A scripting language that is part of the WAP specification.

Distributed Approach for QoS Guarantee to Wireless Multimedia

Kumar S. Chetan
NetDevices India Pvt Ltd, India

P. Venkataram
Indian Institute of Science, India

Ranapratap Sircar
Wipro Technologies, India

INTRODUCTION

Providing support for QoS at the MAC layer in the IEEE 802.11 is one of the very active research areas. There are various methods that are being worked out to achieve QoS at MAC level. In this article we describe a proposed enhancement to the DCF (distributed coordination function) access method to provide QoS guarantee for wireless multimedia applications.

Wireless Multimedia Applications

With the advancement in wireless communication networks and portable computing technologies, the transport of real-time multimedia traffic over the wireless channel provides new services to the users. Transport is challenging due to the severe resource constraints of the wireless link and mobility. Key characteristics of multimedia-type application service are that they require different quality of service (QoS) guarantees.

The following characteristics of WLAN add to the design challenge:

- Low bandwidth of a few Mbps compared to wired LANs bandwidth of tens or hundreds Mbps.
- Communication range is limited to a few hundred feet.
- Noisy environment that leads to high probability of message loss.
- Co-existence with other potential WLANs competing on the same communication channel.

Successful launching of multimedia applications requires satisfying the application's QoS requirements.

The main metrics (or constraints) mentioned in such guidelines and that eventually influence the MAC design are: time delay, time delay variation and data rate. We develop a scheme to provide guaranteed data rate for different applications in WLAN environment.

PROPOSED ENHANCEMENT OF DCF TO PROVIDE QoS

The proposed enhancement is developed as a modular system, which integrates with DCF MAC of 802.11b wireless LAN.

Salient Features of the Modular System

- Provides throughput guarantee for traffic flow between a pair of mobile stations.
- Works in distributed mode.
- Provides MAC level admission control for traffic flow.
- Applications on the mobile stations can send resource reservations request for each call (session).
- Works with backward compatibility, on hosts that do not support QoS enhancements.

Based on the basic principle of DCF access mode, each mobile station transmits data independent of other mobile station. Also, the AP (access point) has no role to play during the data transmission. Under this scenario, the throughput control (and guarantee) has to be achieved in a distributed manner.

One has to restrain a station from accessing the medium if there are other stations in the BSS that has requested for higher resource. If there is no such other station, the station is allowed to access the medium. We propose to use eight different priority flows. The queue manager at mobile stations maintains queues for these flows. Also, the state of these queues (if there are applications that are using this flow) is synchronized across all the mobile station via the beacon messages sent by AP. The scheduler transmits the

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Block schematic of proposed system at AP

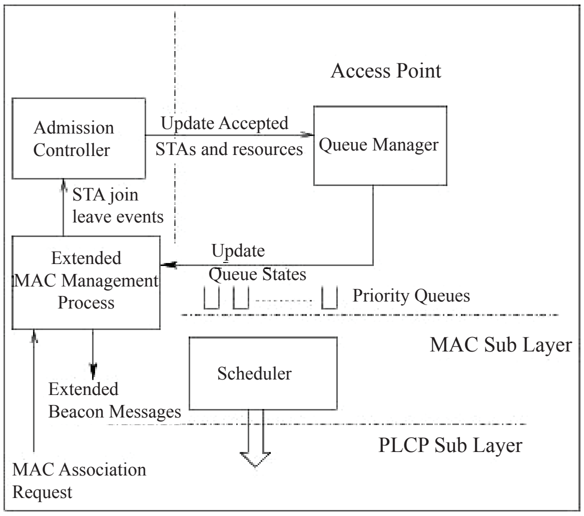
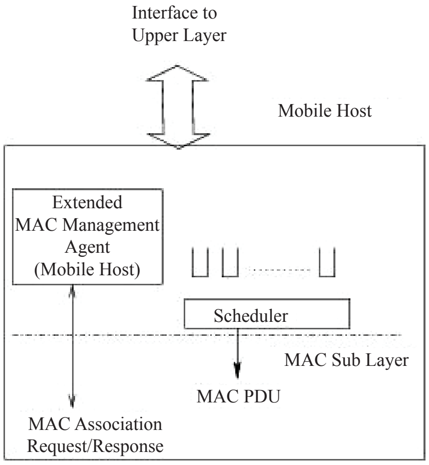

Access Point
Admission Controller
Update Accepted STAs and resources
Queue Manager
STA join leave events
Extended MAC Management Process
Update Queue States
Priority Queues
MAC Sub Layer
Scheduler
Extended Beacon Messages
PLCP Sub Layer
MAC Association Request

Figure 2. Block schematic of proposed system at mobile station


Interface to Upper Layer
Mobile Host
Extended MAC Management Agent (Mobile Host)
Scheduler
MAC Sub Layer
MAC PDU
MAC Association Request/Response

packets from these priority queues using a priority algorithm. The admission controller admits a call to particular flow, if the acceptance of call to the flow do not over-shoot the throughput for that flow.

ARCHITECTURE OF THE MODULAR SYSTEM

The block schematics of the proposed system are given in Figures 1 and 2, for AP and mobile station respectively. The system has four major components:

1. Extended MAC management process
2. Admission controller
3. Queue manager
4. Scheduler

The detailed functioning of each of the components is explained in the following subsections.

Enhanced MAC Management Process

To signal the QoS messages, we propose an extension to MAC layer management messages to carry the resource request and responses.

To signal mobile host resource requirements to AP, we propose extension to the existing MAC management frames. This approach does not need any changes in the core MAC layer. The SME (station management entity), which is normally residing in a separate management plane, needs modifications, which can be easily incorporated.

We incorporate enhanced MAC management with two segments.

MAC Management Process at AP

Apart from receiving and transmitting extended management frames to signal QoS, here the AP also broadcasts the queue states as an extension to beacon message.

MAC Management agent at Mobile Host

Apart from normal functionalities, the agent also receives the extended beacon message with queue states and passes this information to QM.

Management Message Modified

The beacon message is extended to include the queue states in a bit mask format. The queue state is an eight-bit field with each field representing a priority queue state. The queue is considered active if the bit is set, else inactive.

Admission Controller

The admission controller gets triggered by the extended MAC management process. The admission controller is a parameter based admission controller. However the decision process is modified to suit the distributed nature of DCF functionalities. While working in DCF mode each mobile station transmits/receives PDUs independent of AP and independent of other mobile stations. However the queue

Figure 3. Queue state field

Priority Queue 1	Priority Queue 2	Priority Queue 3	Priority Queue 4	Priority Queue 5	Priority Queue 6	Priority Queue 7	Priority Queue 8
bit 1							bit 8

Bit = 1 Queue is active
Bit = 0 Queue is inactive

Algorithm 1. Call admission controller algorithm

```
Begin
    for ( i = Min_Priority;i <= Max_Priority; i++ )
        do
            if ( Requested_Tput + Current_Tput[i] <= Max_Tput[i] )
                then
                    goto accept_call
                else
                    continue
            endif
        done
    Call RejectCall() /* No priority flow could fit this call */
    goto end

    accept_call:
        Current_Tput[i] = Current_Tput[i] + Request_Tput
        Call AcceptCall()
End
```

manager and scheduler as designed such that each station (while being completely independent) will transmit the PDUs at certain priority class. In order for the admission controller to admit a call, it needs to identify the maximum available throughput for a particular flow.

Based on [3] and [4] maximum achievable throughput per priority class is estimated as follows.

$$\text{Max Tput}[i] = \frac{P(\text{successful tx} \mid \text{flow} = i) * \text{Data pay load size}}{P(\text{collision}) * \text{Dur}_{\text{collision}} + P(\text{slot is idle}) * \text{aSlotTime} + P(\text{successful tx}) * \text{Dur}_{\text{success}}} \quad (1)$$

where $\text{Dur}_{\text{success}}$ is time duration for successful transmission of PDU, $\text{Dur}_{\text{collision}}$ is collision duration and aSlotTime is duration of one slot interval.

When a new call request arrives, the call is tried to fit in a flow of highest priority. By doing so, if the call requested throughput is achieved and existing calls are not disturbed, the call is accepted, or else the priority is decreased till the call is acceptable, if the call is not acceptable beyond least priority, the call is rejected. The admission control algorithm is described as follows.

Queue Manager

The main functionality of the queue manager is to synchronize the queue states across all mobile stations. Due to lack of any centralized coordinator, the queue manager synchronizes the states using broadcast messages. The queue manager at the AP broadcasts the states of queues to all the mobile stations in the BSS using extension to Beacon frames.

A queue for a particular flow is active if there are any calls in that flow, otherwise the queue is inactive. The queue states are broadcasted using bit masks in the extended beacon message, with each bit representing a flow. If the bit is set to 1, the queue for that flow is active; otherwise the queue is inactive. The queue manager module on station receives these broadcast messages and updates the queue state.

When all the stations have exchanged the states of the queue, each of the mobile stations would have synchronized queue states. Thus all the stations in the BSS will have a queue at particular priority level as active.

Scheduler

The scheduler's job to pick a packet from the priority flow queue and schedule them for transmission. The scheduler

Algorithm 2. Scheduling packets

```
begin
        r  = rand()    /* r[0,1] */
        CurrentProb   0
        for(i=1 to N) /* N is number of active flows */
                do
                        if (r < CurrentProb)
                                Call TransmitPacket(i)
                        else
                                CurrentProb  = CurrentProb+NormPriority[i]
                        endif
                done
end
```

picks the packet from active queues. Irrespective of the fact the queue has data or not, the PacketTransmit function is called by the scheduler. The scheduler picks up the packet randomly from any of the active queues only. Now the probability to choose a queue is equal to normalized value of the priority of the queue. We define the normalized priority P_n as

$$P_{n_i} = \frac{P_i}{\sum_K P_k} \tag{2}$$

where P_i is the priority of i^{th} queue.

We have assumed equal sized packets in all queues. For non-equal packet sizes the normalized priority would become:

$$P_{n_i} = \frac{\frac{P_i}{S_i}}{\sum_k \frac{P_k}{S_k}} \tag{3}$$

where S_k is the size of k_{th} packet.

Each of the active queues is arranged in the ascending order of the priority. The scheduler selects the queue to be scheduled, and send the packet for transmission. The scheduling algorithm is described in Algorithm 2.

Packet Transmission Process

The PacketTransmit function checks for the packet and hands over the packet to MAC layer if the queue has data.

If the queue is empty, no packet is passed to the MAC function; however the packet transmission will back-off for other stations to transmit. A timer call back for a duration that represents the time taken by an average-sized packet is registered. This will provide opportunity for other stations with the packet in the same priority level queue to transmit the packet on to the medium.

The MAC will perform DCF access method to access the medium and make transmission. If the packet could not be transmitted due to wireless medium loss or collusion, the packet is retransmitted by the MAC layer, the scheduling is not changed due to retransmission. The packet transmission algorithm is described in Algorithm 3.

FUNCTIONING OF THE MODULAR SYSTEM

The functions of modular system are to gather the QoS requirements from the mobile stations and schedule their applications as per the specified QoS.

The functioning of the system is explained by considering the following cases.

Case 1. When New Mobile Station Enters the BSS with New Application

When a new station enters the BSS, the station joins the BSS using normal association procedure (as per IEEE802.11 standard). The AP instantiates a MAC management agent. The agent is migrated to the mobile host. The station uses the "Extended MAC Management Agent" to signal the re-

Algorithm 3. Packet transmission algorithm

```
Begin
        if ( i == active) /* Is the queue non-empty */
                Call SendPacket /* Call the MAC function */
                Call TimerCallback(packet_size)
        else
                Call TimerCallback(Ave_packet_size)
        endif
end
```

source requirements for new application. The AP receives the resource request, and passes on the QoS parameters to admission controller. The admission controller examines the resource requests, applies the admission control algorithm and accepts/rejects the request. The accept/reject response is conveyed back to the mobile station via the "Extended MAC Management process" module and updates the queue states in the QM. The QM updates the queue states for each of the flows, and sends them to the mobile stations as an extension to beacon message.

Upon reception of the response for the call request, the mobile station can start sending traffic. The QM at the mobile station receives the queue states in the extended beacon message and updates local queue state. The scheduler the mobile station schedules the traffic using the queue state information.

Case 2. When a Mobile Station Leaves the BSS

When the mobile station is about leave the BSS, the station uses the enhanced MAC management process to signal the AP. The AP receives the station ID, and updates the QM and admission control to relinquish the resource used all application under this station and changes the queue state if required.

Case 3. When New Application Has to be Admitted

A mobile station is already associated with the AP, and may be running some applications or otherwise. Under this condition, the station and AP follow the normal steps (as in Case 1), as this is no different from Case 1.

Case 4. When Applications are Terminated at the Mobile Station

A mobile station stops the current application, but the station is still in the BSS. When the application is stopped, the station uses the enhanced MAC management process to signal the AP that the application has stopped. The AP follows steps in Case 2 to relinquish the resource and update queue state.

IMPLEMENTATION CONSIDERATION

The system has two components, that is to say, AP component and mobile station component.

- The AC is implemented at the centralized location, normally co-located with the AP.
- Extension to MAC management layer, QM and scheduler are implemented in both AP and mobile station.

AP Components Implementation

It is easier to enhance the AP to support functionalities such as CAC, QM and scheduler since there is relatively very few numbers of APs and also the APs are normally controlled by the service provider, who wishes to provide QoS in the BSS. The AC, QM and scheduler are implemented in high-level system language (C) and integrated into the AP software. The MLME functions on the AP are modified to understand the extensions in the MAC frames and pass the QoS parameters to the CAC. The receive (RX) and transmit (TX) functions on the AP are modified to call the QM and scheduler functions respectively.

Mobile Station Components Implementation

The station components are implemented as Java byte-code, and the agent is instantiated at the AP and migrated to the mobile station, as the mobile station gets associated with the AP. The extended MAC management agent, QM and scheduler are implemented in Java.

SIMULATION AND RESULTS

We have used ns-2 [7] as simulator to validate the proposed system. Ns are a discrete event simulator targeted at networking research. Ns provides substantial support for simulation of TCP, routing, and multicast protocols over wired and wireless (local and satellite) networks.

Simulation Setup and Procedure

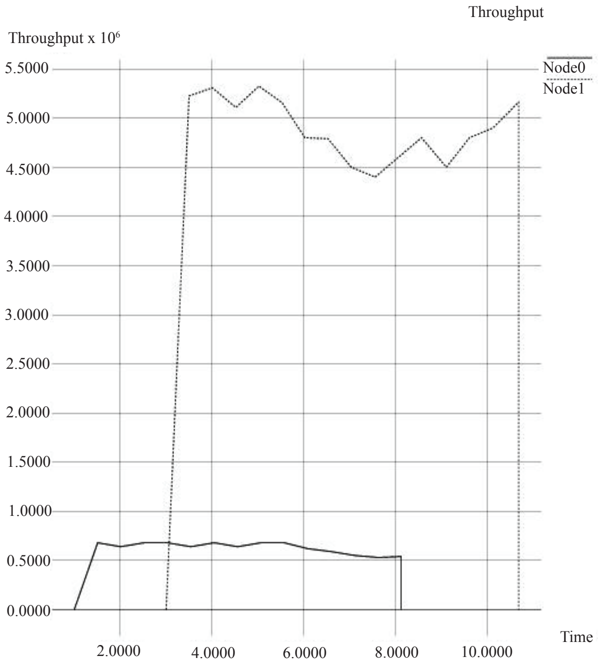
Using ns-2, a sixteen-node network working in BSS and iBSS modes is setup. We have modified the MAC layer functionalities in ns-2 to support the proposed system.

In the simulated system, the mobile nodes communicate with each other and external world over the 802.11b 11MBPS channel. The transmission power of the mobile is chosen such that stations are not hidden from one another. If not stated otherwise, during the simulation neither RTS/CTS nor the fragmentation is used. For most of our simulation we have considered on-off traffic with exponential distribution, and log-normal distribution for the radio channel errors. We have done several experiments by considering traffic in both upstream and downstream direction. The allowed traffic has been classified as VBR (variable bit-rate) and

Table 1. Simulation parameters

Data Rate	11 Mbps
RTS Threshold	3000
Physical Layer Frequency	2457e+6
Transmission Power	31 mW
Receiver Threshold	1.15209e-10
Radio Propagation Model	TwoRayGround
LongRetryLimit	2

Figure 4. Throughput plot for mobile station with CBR traffic under DCF mode



CBR (constant bit-rate) with random arrivals. The duration of each experiment varies from 100 seconds to 1,000 seconds. The simulation parameters that are significant are listed in Table 1.

We have simulated the proposed systems as three individual units. We explain the simulation for each case with CBR and VBR traffics.

The DCF MAC functions at each station are modified according to the proposed scheme. Each of the mobile station generates traffic destined to other nodes with data packets of 1,500bytes, 1,000 bytes, 512 bytes and 64 bytes.

In our first experiment, we considered two nodes generating CBR traffic. Each node requested a throughput of 10\% and 90\% of the net throughput. In Figure 4, we have plotted the throughput vs. time for two nodes. It can be observed that both the stations get the allocated bandwidth.

Figure 5. Throughput plot for mobile station with VBR traffic under DCF mode

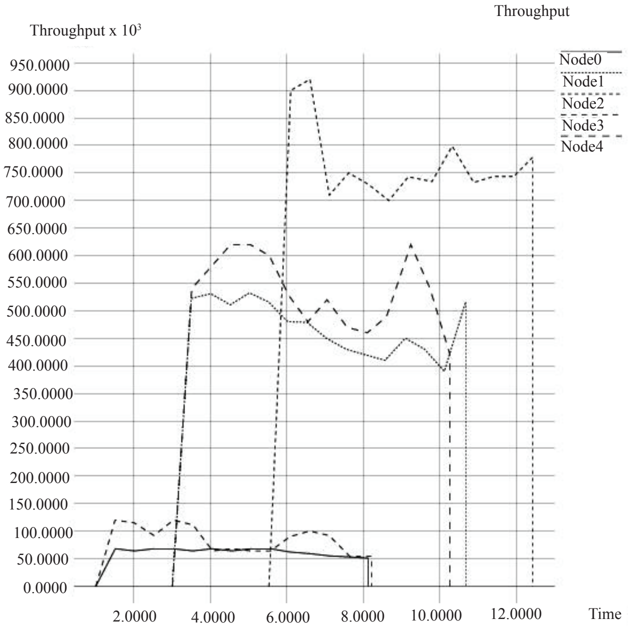
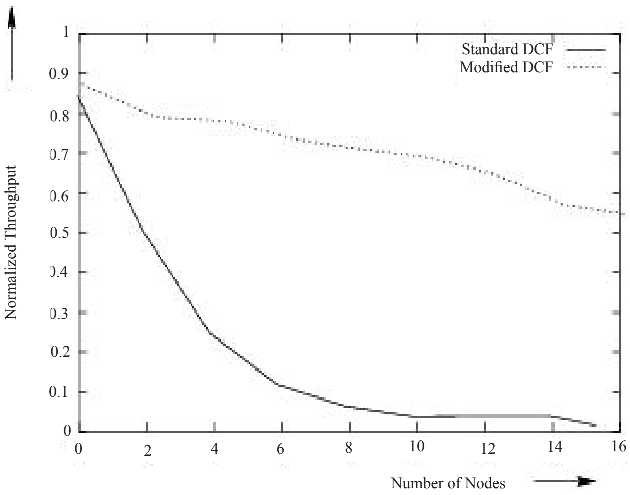


Figure 6. Normalized throughput for proposed scheme and standard DCF mod



Also when the node0 application is terminated, the node1 uses up the excess bandwidth.

In the next experiment, we have considered BSS with five nodes admitted for scheduling. The nodes requested throughput requirements are at 50, 100, 500, 550, and 750 Kbps. In Figure 5, we have plotted the throughput versus time for all the five nodes. It can be observed from the plot, each of the nodes is scheduled to get the allocated bandwidth.

Comparison with the Standard DCF Scheme

Again we use the normalized throughput as a measure to compare the proposed DCF scheme against the standard DCF scheme.

We have plotted the normalized throughput for a node working proposed DCF mode (with CBR traffic) against standard DCF mode in Figure 6. With the proposed scheme, the mobile 6station is admitted to the BSS via admission controller and the scheduler schedules the allocated bandwidth for the station. Hence the normalized throughput stays close to one. With the standard DCF, the mobile stations share the bandwidth in uncontrolled manner. Thus, as the number of stations increases in the BSS, there is no control over the bandwidth usage.

SUMMARY

In this article we have proposed method of providing QoS in wireless LAN operating in DCF mode. Providing QoS guarantee in a distributed environment has to be a distributed approach due to the distributed nature of the DCF. The proposed method requires small changes (extensions) to the MAC management entity and scheduler function operating at each station above the MAC layer.

REFERENCES

Anker, T., Cohen, R., Dolev, D., & Singer, Y. (2001). Probabilistic fair queuing. In *IEEE Workshop on High Performance Switching and Routing.* Dallas, TX.

Bianchi, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE JSAC,* 3, 535-47.

Chetan Kumar, S., Venkataram, P., & Pratap Singh, R. (2005). Distributed approach for QoS guarantee to wireless multimedia. In *International Conference on Advances in Mobile Multimedia,* Kuala Lumpur, Malaysia.

Network Simulator NS-2. (n.d.). Retrieved from http://www.isi.edu/nsnam/ns/.

Ni, Q., Romdhani, L., Turletti, T., & Aad, I. (2002). QoS issues and enhancements for IEEE 802.11 wireless LAN. In *Rapport de recherche de l INRIA.* Sophia Antipolis, Equipe: PLANETE.

Pong, D., & Moors, T. (2003). Call admission control for IEEE 802.11 contention access mechanism. *Globecom 2003,* San Francisco USA.

Tay, Y. C., & Chua, K.C. (2001). A capacity analysis for the IEEE 802.11 MAC protocol. *Journal of Wireless Networks, 7.*

KEY TERMS

Access Point (AP): An entity in the wireless LAN that is normally connected to a wired backbone and coordinates the operation of wireless mobile stations that are operating in infrastructure mode.

Call Admission Controller (CAC): A set of functions and procedures that control the admission of new calls into the system predictable based on predefined set of rules.

Media Access Controller (MAC): A set of procedures that governs the access to the media in a multiple access system. DCF (distributed coordination function) and PCF (point coordination function) are two MAC functions defined in IEEE 802.11 standard for wireless LAN.

Quality of Service (QoS): The term quality of service defines of quantitative representation of network resources that affect the application performance.

Distributed Computing in Wireless Sensor Networks

Hong Huang
New Mexico State University, USA

INTRODUCTION

Wireless sensor networks (WSNs) recently have attracted a great amount of attention because of their potential to dramatically change how humans interact with the physical world (Estrin, Culler, Pister, & Sukhatme, 2002; Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002). A wireless sensor network is composed of many tiny, wirelessly connected devices, which observe and perhaps interact with the physical world. The applications of WSN are many and wide-ranging, including wildlife habitat monitoring, smart home and building, quality monitoring in manufacturing, target tracking in battlefields, detection of biochemical agents, and so forth.

The emerging WSN technology promises to fundamentally change the way humans observe and interact with the physical world. To realize such a vision, *distributed computing* is necessary for at least two reasons. First, sending all the raw data to a base station for centralized processing is very costly in terms of energy consumption and often impractical for large networks because of the scalability problem of wireless networks' transport capacity (Gupta & Kumar, 2000). Second, merely using a large number of inexpensive devices to collect data hardly fundamentally changes the way humans interact with the physical world; and it is the intelligence embedded inside the network (i.e., distributed computing) that can have a profound impact.

WSN presents a very difficult environment for distributed computing. Sensors have severe limitations in processing power and memory size, and being battery-powered, they are particularly energy constrained. As a reference, the hardware capabilities of a typical sensor node are listed in Table 1. In WSN, device failures can be frequent, sensory data may be corrupted by error, and the wireless communications exhibit complex and unpredictable behavior. In such an environment, traditional methods for distributed computing face fundamental difficulties. Now, communications links are neither reliable nor predictable: they can come and go at any time. Packet routing is difficult since maintaining and storing routing tables for a massive number of nodes is out of the question. Routing to a single destination seems to have a solution (Intanagonwiwat, Govindan, & Estrin, 2000), complex message routing for distributed computing remains difficult. Also, distributed organizing and grouping of sensory data using traditional methods is costly in terms of protocol message overhead.

This article is organized as follows. We first describe WSN infrastructures required to support distributed computing, followed by a description of typical, important distributed computing applications in WSN; we the conclude the article.

INFRASTRUCTURE SUPPORT FOR DISTRIBUTED COMPUTING IN WSN

In order for WSN to effectively perform distributed computing, some necessary infrastructure needs to be established. The type of infrastructure required varies according to the specific application in question, but the common ones include neighbor discovery and management, synchronization, localization, clustering and grouping, and data collection infrastructure. We elaborate on each item as follows.

Table 1. Hardware capabilities of typical sensor nodes (Crossbow Technology —www.xbow.com)

	CPU	Nonvolatile Memory	Radio Transceiver	Power
MICA2	ATMega 128L 8 MHz, 8 bit	512 KB	869/915, 434, 315 MHz, FSK ~40 Kbps	2 AA 2850 mAh
MICA2DOT	ATMega 128L 8 MHz, 8 bit	512 KB	869/915, 434, 315 MHz, FSK ~40 Kbps	Coin cell 1000 mAh

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Neighbor discovery and management refers to the process in which sensors discover their neighbors, learn their properties, and control which neighbors to communicate with. Discovery is typically done through sensors exchanging hello messages within radio range. In the process, sensors discover not only neighbors' presence, but also optionally their node type, node identifier, power level, location/coordinates, and so forth. Frequently, sensors can also control how many neighbors to communicate with through the use of power control—that is, a sensor can increase or decrease the scope of its immediate neighborhood by increasing or decreasing its transmitting power, respectively. This is also called topology control (Li & Hou, 2004), and its purpose is to allow sensors to use just enough, but no more power to ensure adequate connectivity.

Synchronization refers to the process in which sensors synchronize their clocks. Synchronization is necessary because sensory data is often not useful if not put in a proper temporal reference frame. Traditional methods for synchronization in a network, such as NTP (Mills, 1994), do not apply very well in WSN. This is because the assumptions on which tradition network synchronization methods are based, such as availability of high-precision clocks, stable connections, and consistent delays, are no longer true in WSN, causing considerable difficulties. The approach to deal with such difficulties in WSN is to relax the requirements. For example, only local, not global, synchronization is maintained, or only event ordering, not precise timing, is kept (Romer, 2001).

Localization refers to the process in which sensors obtain their position/coordinates information. Similar to synchronization, localization is necessary because sensory data needs to be put in a spatial reference frame. Sensors with global positioning system (GPS) capability are currently commercially available; they obtain their coordinates from satellites with a few meters' accuracy. The downside with using GPS is the cost, and the unavailability indoors or under dense foliage. In a WSN without GPS capability, it is still possible to localize relatively to a few reference points in the network (Bulusu, Heidemann, & Estrin, 2000).

Clustering and grouping refers to the process in which sensors organize themselves into clusters or groups for some specific function. A cluster or a group typically consists of a leader and a few members. The leader represents the cluster or group and maintains external communication, while the members report data to the leader and do not communicate with the outside. Such organization is advantageous for scalability, since a large network can now be reduced to a set of cluster or groups. Task-specific clusters or groups can be formed. For example, sensors around a moving target form a tracking group, which moves with target, while sensors not in the tracking group can be put to sleep to save energy (Liu, Reich, Cheung, & Zhao, 2003).

Data collecting infrastructure ensures that sensory data is transported correctly and efficiently to one or a few collection points, sometimes called data sinks. A typical approach is publish and subscribe with attribute-based naming, where a sink broadcasts its interest for some data attributes, and sensors send their data if it matches the interests. An example of such an approach is directed diffusion, in which an infrastructure based on the hop count to the sink is established and refreshed periodically (Intanagonwiwat et al., 2000).

TYPICAL DISTRIBUTED COMPUTING APPLICATIONS IN WSN

In this section, we describe typical distributed computing applications in WSN which include distributed query and search, collaborative signal processing, distributed detection and estimation, and distributed target tracking.

Distributed query and search refers to the process in which a user query or search for an event or events inside the network in a distributed fashion. There are two major types of such applications: blind and structured. In a blind search, no prior information about the target exists. In a structured search, some kind of infrastructure exists which points to the target location in a distributed manner. We elaborate on these two types of searches below.

There are three major approaches to perform blind search. The first one is flooding, in which the query message is flooded to the entire network and the target responds with a reply. The advantages of flooding are simplicity and low response latency. The disadvantage is the high communications cost in terms of number of messages transmitted. To mitigate the high communication cost of flooding, a second approach, iterative, limited flooding, can be used (Chang & Liu, 2004). In such an approach, a sequence of limited broadcasts of increasing hop-count limits is tried until the target is found, in the hope that the target will be found during a low-cost, limited broadcast. The expected communications cost reduction of this approach comes at the expense of higher search latency. In the third approach, a query packet carries out a random walk in the network, which continues until the search target is encountered (Avin & Brito, 2004). This approach can further reduce communications cost but at the expense of even higher latency.

In a structured search, indices or pointers for targets are distributed in the network. A typical approach uses a distributed hash table, where the name of a target is randomly and uniformly hashed to a number that identifies a node, or a location, where the target information is stored (Ratnasamy et al., 2002). The search becomes a simple matter of evaluating the hash function of the target name which points to a node that stores the target information. This simplification comes at the cost of maintaining an infrastructure that stores target information in a distributed manner, which can be costly if the network is highly dynamic, for example, nodes join or leave the network frequently.

Collaborative signal processing (CSP) refers to the process in which sensors process sensory data collaboratively rather than individually. CSP has three main advantages. First, CSP promotes robustness because, while an individual sensor can be faulty, consensus from many sensors is reliable. Second, CSP promotes accuracy since the results fuse the observations from many sensors, canceling noise from individual sensors. Third, CSP promotes efficiency since only a fraction of sensors needs to be activated—those which have highest potential to eliminate uncertainty (Zhao, Shin, & Reich, 2002). In addition, sophisticated multi-resolution analysis techniques, such as wavelet transform, can be used to reduce the amount of data transported. CSP is a common method used in different distributed applications, and its specifics vary according to the particular application in question and are described in the following sections.

Distributed detection and estimation refers to the process in which WSN makes a decision about the occurrence of an event (detection) or about the value of a physical variable (estimation) in a distributed manner. In distributed detection, under both Bayes and Neyman-Peason criterion, the optimum test is a likelihood ratio test. In the case of independent observations, it suffices to send likelihood ratios from individual sensors rather than the raw data, leading to a large savings in communications cost. In distributed estimation, the principal of sufficient statistics can be appealed to drastically reduce the amount of data transported. For example, to estimate the mean of a Gaussian variable, the sufficient statistic is simply the sum of sampled values and the count of number of samples, both of which can be collected using a running sum or count rather than the entirety of the sampled values, resulting in large savings in communications cost.

Research issues related to distributed detection and estimation include sensor data censoring to exclude spurious data (Patwari & Hero, 2003), distributed, iterative sensor fusion based on consensus (Xiao, Boyd, & Lall, 2005), and finding the optimal tradeoff between decision fidelity or estimation error, and sensor density, sensitivity, quantization level, communications cost, and power consumption (Aldosari & Moura, 2004).

Distributed target tracking refers to the process in which sensors collaboratively track one or more discrete, moving targets. In a sense, distributed target tracking is a subclass of distributed detection and estimation, but faces the severe challenge caused by target movement. To deal with such difficulty, sensors typically form a tracking group (Li, Wong, Hu, & Sayeed, 2002). The group membership can be determined in a number of ways. A sensor can declare itself a member if it receives a signal from the target that exceeds a certain threshold. Or sensors can be added to the group sequentially in the order of its capability to reduce uncertainty (Zhao et al., 2002). The sensory data from members of the tracking group is collected at the group leader, which can be selected as the sensor with the highest signal level. For the fusion of sensory data, a number of standard methods are available, such as Kalman filtering (Spanos, Olfati-Saber, & Murray, 2005; Mokashi, Huang, Kuppireddy, & Varghese, 2005), maximum likelihood estimation, and so forth.

Research issues related to distributed target tracking include the distribution of state (target) information among individual sensors (Liu, Chu, Reich, & Zhao, 2004), managing tracking groups, and trading off power consumption and surveillance quality (Gui & Mohapatra, 2004).

CONCLUSION

In this article we have provided an introduction to WSN and its characteristics, and the challenges WSN poses for distributed computing. We have described some infrastructure support for distributed computing, including neighbor discovery and management, synchronization, localization, clustering and grouping, and data collection. We have discussed typical, important distributed computing applications in WSN, including distributed query and search, collaborative signal processing, distributed detection and estimation, and distributed target tracking.

WSN presents new, exciting challenges to distributed computing, such as how to achieve complex, high-level goals from a large number of simple devices, and how to obtain robustness and certainty through coordinating unreliable devices and processing uncertain data. Successfully addressing such challenges will significantly advance the state of the art of distributed computing.

REFERENCES

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). A survey on sensor networks. *Computer Networks, 38*(4), 393-422.

Aldosari, S. A., & Moura, J. M. F. (2004). Fusion in sensor networks with communication constraints. *Proceedings of the Conference on Information Processing in Sensor Networks (IPSN'04)*, Berkeley, CA.

Avin, C., & Brito, C. (2004). Efficient and robust query processing in dynamic environments using random walk techniques. *Proceedings of the Conference on Information Processing in Sensor Networks (IPSN'04)*, Berkeley, CA.

Bulusu, N., Heidemann, J., & Estrin, D. (2000). GPS-less low-cost outdoor localization for very small devices. *IEEE Personal Communications, 7*(4), 28-34.

Chang, N., & Liu, M. (2004). Revisiting the TTL-based controlled flooding search: Optimality and randomization. *Proceedings of the International Conference on Mobile Computing and Networks (MobiCom'04)*, Philadelphia, PA.

Estrin, D., Culler, D., Pister, K., & Sukhatme, G. (2002). Connecting the physical world with pervasive networks. *IEEE Pervasive Computing*, *1*(1), 59-69.

Gui, C., & Mohapatra, P. (2004). Power conservation and quality of surveillance in target tracking sensor networks. *Proceedings of the International Conference on Mobile Computing and Networks (MobiCom'04)*, Philadelphia, PA.

Gupta, P., & Kumar, P. R. (2000). The capacity of wireless networks. *IEEE Transactions on Information Theory, 46*(2), 388-404.

Intanagonwiwat, C., Govindan, R., & Estrin, D. (2000). Directed diffusion: A scalable and robust communication paradigm for sensor networks. *Proceedings of the International Conference on Mobile Computing and Networks (MobiCom'00)*, Boston.

Li, D., Wong, K., Hu, Y., & Sayeed, A. (2002). Detection, classification, tracking of targets in micro-sensor networks. *IEEE Signal Processing Magazine, 19*(2), 17-30.

Li, N., & Hou, J. C. (2004). Topology control in heterogeneous wireless networks: Problems and solutions. *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'04)*, Hong Kong, China.

Liu, J., Reich, J., Cheung, P., & Zhao, F. (2003). Distributed group management for track initiation and maintenance in target localization applications. *Proceedings of the Conference on Information Processing in Sensor Networks (IPSN'03)*, Palo Alto, CA.

Liu, J., Chu, M., Reich, J., & Zhao, F. (2004). Distributed state representation for tracking problems in sensor networks. *Proceedings of the Conference on Information Processing in Sensor Networks (IPSN'04)*, Berkeley, CA.

Mills, D. L. (1994). Precision synchronization of computer network clocks. *ACM Computer Communication Review*, *24*(2), 28-43.

Mokashi, G., Huang, H., Kuppireddy, B., & Varghese, S. (2005). A robust scheme to track moving targets in sensor nets using amorphous clustering and Kalman filtering. *Proceedings of the IEEE Military Communications Conference (Milcom'05)*, Atlantic City, NJ.

Patwari, N., & Hero, A. O. (2003). Hierarchical censoring for distributed detection in wireless sensor networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, Hong Kong, China.

Ratnasamy, S., Karp, B., Yin, L., Yu, F., Estrin, D., Govindan, R., & Shenker, S. (2002). GHT: A geographic hash table for data-centric storage. *Proceedings of the ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02)*, Atlanta, GA.

Romer, K. (2001). Time synchronization in ad hoc networks. *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'01)*, Long Beach, CA.

Spanos, D., Olfati-Saber, R., & Murray, R. (2005). Distributed Kalman filtering in sensor networks with quantifiable performance. *Proceedings of the Conference on Information Processing in Sensor Networks (IPSN'05)*, Los Angeles, CA.

Xiao, L., Boyd, S., & Lall, S. (2005). A scheme for asynchronous distributed sensor fusion based on average consensus. *Proceedings of the Conference on Information Processing in Sensor Networks (IPSN'05)*, Los Angeles, CA.

Zhao, F., Shin, J., & Reich, J. (2002). Information-driven dynamic sensor collaboration for tracking applications. *IEEE Signal Processing Magazine, 19*(2), 61-72.

KEY TERMS

Clustering and Grouping: The process in which sensors organize themselves into clusters or groups for some specific function.

Collaborative Signal Processing (CSP): The process in which sensors process sensory data collaboratively rather than individually.

Data Collecting Infrastructure: Ensures that sensory data is transported correctly and efficiently to one or a few collection points.

Distributed Computing: The kind of computing realized through, and distributed among, many discrete devices.

Distributed Detection and Estimation: The process in which a WSN makes a decision about the occurrence of an event (detection) or about the value of a physical variable (estimation) in a distributed manner.

Distributed Query and Search: The process in which a user queries or searches for an event or events inside the network in a distributed fashion.

Distributed Target Tracking: The process in which sensors collaboratively track one or more discrete, moving targets.

Localization: The process in which sensors obtain their position/coordinates information.

Neighbor Discovery and Management: The process in which sensors discover their neighbors, learn their properties, and control which neighbors to communicate with.

Synchronization: The process in which sensors synchronize their clocks.

Wireless Sensor Network (WSN): A type of network composed of many tiny, wirelessly connected sensing devices.

Distributed Heterogeneous Tracking for Augmented Reality

Mihran Tuceryan
Indiana University Purdue University Indianapolis, USA

Rajeev R. Raje
Indiana University Purdue University Indianapolis, USA

INTRODUCTION

Augmented reality (AR) is a technique in which a user's view of the real world is enhanced or augmented with additional information generated from a computer model (Azuma et al., 2001). The enhancement may consist of virtual artifacts to be fitted into the environment or a display of non-geometric information about existing real objects. Mobile AR (MAR) systems implement this interaction paradigm in an environment in which the user moves, possibly over wide areas (Feiner, MacIntyre, Hoellerer, & Webster, 1997). This is in contrast to non-mobile AR systems that are utilized in limited spaces such as a computer-aided surgery or by a technician's aid in a repair shop. There are a number of challenges to implementing successful AR systems. These include a proper calibration of the optical properties of cameras and display systems (Tuceryan et al., 1995; Tuceryan, Genc, & Navab, 2002), and an accurate registration of three-dimensional objects with their physical counterparts and environments (Breen, Whitaker, Rose, & Tuceryan, 1996; Whitaker, Crampton, Breen, Tuceryan, & Rose, 1995). In particular, as the observer (or an object of interest) moves over time, the 3D graphics need to be properly updated so that the realism of the resulting scene and/or alignment of necessary objects and graphics are maintained. Furthermore, this has to be done in real time and with high accuracy. The technology that allows this real-time update of the graphics as users and objects move is a *tracking system* that measures the position and orientation of the tracked objects (Koller et al., 1997). The ability to track objects, therefore, is one of the big challenges in MAR systems. This article describes a software framework for realizing such a distributed tracking environment by discovering independently deployed, possibly heterogeneous trackers and fusing the data from them while roaming over a wide area. In addition to the MAR domain, this kind of a tracking capability would also be useful in other domains such as robotics and location-aware applications. The novelty of this research lies in the amalgamation of the theoretical principles from the domains of AR/VR, data fusion, and the distributed software systems to create a sensor-based, wide-area tracking environment.

BACKGROUND

Although a few approaches for tracking have been proposed (e.g., Hightower & Borriello, 2001; Koller et al., 1997; Neumann & Cho, 1996; State, Hirota, Chen, Garrett, & Livingston, 1996), the ability to track objects accurately and in real time over a *wide area* does not yet have a satisfactory solution. Moreover, many of these approaches require a highly engineered environment with a uniform set of trackers whose architecture is known in advance (Welch, & Foxlin, 2002; Ubisense, 2006). Assuming that trackers have been deployed and are operating and exist in the environment, this research deals with questions of how to discover what trackers exist in a local area, what quality-of-service (QoS) properties they have, and how to make the best use of their measurements in a mobile and dynamic environment.

The wide-area, ubiquitous tracking that is the focus of this article has been addressed mainly in the pervasive/ubiquitous computing community. An early tracking system was HiBall that utilized a ceiling instrumented by LED lights (Welch, 1999). The HiBall tracker covered a wider area than a typical magnetic tracking system, and in the implementation its range covered a room or a lab. The scalability of such a system was limited because of the increased cost of extending beyond the size of a lab. The BAT system, which used ultrasound as the core technology (Harter, Hopper, Steggles, Ward, & Webster, 2002; Newman, Ingram, & Hopper, 2001), had a limited resolution. The location sensing system, by Ubisense (2006) uses the ultra-wide-band technology and has a better resolution (6 inches positional accuracy, according to company Web sites).

Researchers at Intel Research studied the use of existing wireless hotspots and cell phone towers to compute location information over wide-areas (Schilit et al., 2003; Borriello, Chalmers, LaMarca, & Nixon, 2005). Bahl and others studied localization techniques using existing Wi-Fi wireless hubs (Bahl & Padmanabhan, 2000; Balachandran, Voelker, & Bahl, 2003). Their methods assume a ubiquitous infrastructure that exists for other purposes (networking) that can be tapped into for localization of users. Typically, their resolution tends to be low and is not sufficient for typical AR applications.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Ubiquitous tracking systems specifically for AR systems have been also studied by Bauer et al. (2002). Newman et al. (2004), and Reitmayr (2001), and have resulted in prototypical systems, some of which are component based (e.g., DWARF by [Bauer, Bruegge, Klinker, MacWilliams, Reicher, Riss, et al., 2001]).

THE UNIFRAME-BASED MOBILE TRACKING SERVICE FOR AR

As indicated earlier, the distributed tracking system is an example of a heterogeneous, distributed computing system. The overall architecture and various components of the distributed tracker subsystem that is the focus of this article are shown in Figure 1.

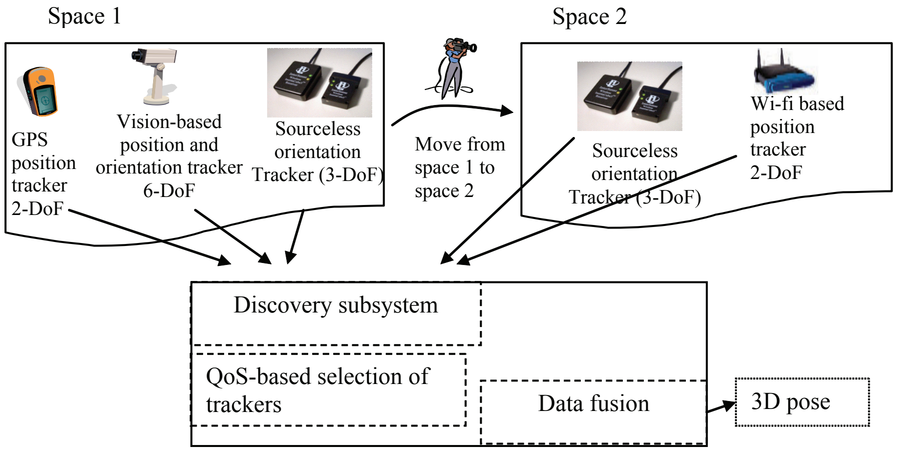
The software realization of this tracking system is based on the principles of uniframe (Olson, Raje, Bryant, Burt, & Auguston, 2005). Uniframe provides an environment for an interoperation of heterogeneous and distributed software components, and uses the principles a meta-component model, service-oriented architectures, generative programming, and two-level (TLG) and event grammars (EG). The realization of the distributed tracking system, using the UniFrame, begins with a generative domain model (GDM) (Czarnecki & Eisenecker, 2000) created by experts from the tracking domain. This GDM contains various details, such as the software architecture of the tracking system expressed in terms of underlying components, their interactions, the rules for generating middleware, and the rules for the prediction and monitoring of the quality of the integrated system. Each component is defined by a Unified Meta-component Model (UMM) specification (Raje, 2000). The UMM has three parts: (a) components; (b) services, offered by components, and associated guarantees; and (c) infrastructure for deploying and discovering components. A developer who wishes to create specific components for the tracking system consults the GDM and creates implementations using the UMM specifications encoded there. After a component is developed, it is validated against the quality requirements, both functional and QoS. The developer also creates an associated UMM specification for that component. This specification and the component are deployed on the network. These components are also registered with the uniframe resource discovery system (URDS) (Siram, 2002).

A system integrator planning to create the tracking system, from independently developed and deployed components, issues a query consisting of the requirements the tracking system must meet. The query processing consults the GDM, divides the query into many sub-queries, each corresponding to a single component UMM specification. These sub-queries are passed to the URDS, which searches for appropriate matching components. If such components are found, these are returned to the system integrator, who then selects a subset of these results, provides any proprietary components, and requests the process to assemble the integrated system conforming to the design. The uniframe system generator (Huang, 2003) carries out the generation of the integrated system. The key challenges in creating the tracking system are: (a) designing the GDM, (b) the discovery of components, and (c) the generation of a prototypical tracking system. These are briefly discussed below.

Designing the GDM

The GDM is developed by the domain experts and contains the software architecture of the family of systems, along with many associated details. For a tracking system, it can be either handcrafted or generated via the uniframe system generator (Huang, 2003). One important piece of the GDM relates to the specification of components that make the software architecture of the tracking system. The specification provides

Figure 1. Architecture and components of the distributed tracking subsystem for mobile AR applications


Space 1
Space 2
GPS position tracker 2-DoF
Vision-based position and orientation tracker 6-DoF
Sourceless orientation Tracker (3-DoF)
Move from space 1 to space 2
Sourceless orientation Tracker (3-DoF)
Wi-fi based position tracker 2-DoF
Discovery subsystem
QoS-based selection of trackers
Data fusion
3D pose

the necessary details during the discovery process and the system generation process. The approach for specifying the components is indicated as follows.

UMM-Specifications

Each sensing device used in the tracking system is represented by a corresponding software component that encapsulates its behavior. For example, a GPS sensor in the tracking system is encapsulated as a component offering a service that provides 3DOF position information with certain accuracy. As indicated earlier, each component in uniframe has an associated UMM specification. This specification contains many attributes (Raje, 2000) that reflect various details related to that component. In the context of the tracking system, the functional and the QoS attributes of a component are the most important ones. For example, a partial specification (for the sake of brevity) for an Inertia Cube Tracker component is:

Component Name: InertiaCubeTracker
Domain Name: Distributed Tracking
Informal Description: Provides the orientation information.
Computational Attributes
Inherent Attributes:
Id: cs.iupui.edu/InertiaCubeTracker;
...
Validity: 12/1/07 Registration: pegasus.cs.iupui.edu/HH1
Functional Attributes:
Functional Description: Provides the orientation of a tracked object.
Algorithm: Kalman Filter;
Complexity: O (n^3)
Syntactical Contract:
Vector getOrientation();
Semantic Contract:
Pre-condition: {calibrated (InertiaCubeTracker)== true}
Post-condition: {sizeof (orientationVector) == 3)}
Synchronization Contract:
Nature: Multi-threaded Synchronization Policy
Implementation Mechanism: semaphore
Technology: CORBA
………
Quality of Service Attributes
QoS Metrics: resolution, drift, lag-time,
resolution: 0.1 degrees; drift: 0.01 degrees/sec; lag-time: 1 ms
………

The preceding partial specification shows many important characteristics of the UMM-specification. These are: (a) the specification is comprehensive and highlights many aspects of a component; (b) the specification is an enhanced realization of the concept of multi-level contracts (Beugnard, Jezequel, Plouzeau, & Watkins, 1999), thus the specification not only describes the functional aspects, but also emphasizes the QoS attributes of a component; and (c) the specification is consistent with the concepts of a service-oriented view of software components. The detailed nature of the specification provides sufficient information for the discovery of components to create the tracking system.

Discovery Process

In a distributed tracking environment, it is conceivable that there would be many different software instances offering similar types of services. For example, there may be multiple trackers (each encapsulating a different inertia tracker) implemented in a variety of ways and possibly offering different qualities of the tracking results. Hence, to realize such a tracking system, it is necessary to discover various alternatives available over the network. That is the role of the discovery service. The discovery process is realized using the principles of the URDS (Siram, 2002). The URDS is a hierarchical, proactive, and interoperable discovery service. The components are selected based on their type and the QoS (e.g., resolution) values for a specific tracking system.

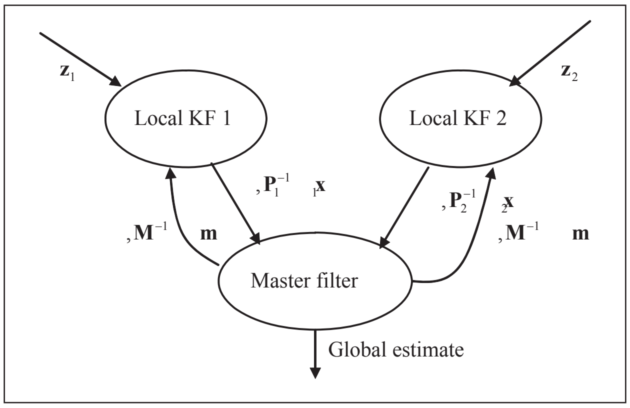
Generation of the Prototypical System

Uniframe uses the principles of glues of wrappers that are generated using the concepts of TLG and EG. The purposes of the glue and wrapper code are to allow an interoperation between heterogeneous components and also to insert any instrumentation code that can collect event traces to observe the actual QoS values during the execution of the system. A system generator (Huang, 2003) accepts the selected components as the input and uses the information in the GDM to semi-automatically create the distributed system. The rules for the generation are a part of the GDM and are developed during the formalization of the GDM. Another important facet of the generation process is the ability to make static predictions (based on the QoS properties of the selected components) and compare them against the actual execution results. The process of prediction is based on the principles of sensor fusion, as described in the following section.

DATA FUSION

Once the individual tracker components are selected through the discovery process described earlier, their results need to be combined (fused) in order to get the best estimate of the position and orientation of the tracked object. The usual

Figure 2. Federated Kalman filter architecture (Carlson, 1990)



Note: Here the $\hat{x}_i$ *and* M_i *are the state vector and covariance matrices for the local filters. The* $\mathbf{z}_i$*s are the measurements for each sensor i. The quantities* $\mathbf{m}$ *and* $\mathbf{M}$ *are the global state vector and covariance matrix estimated by the master filter, respectively. In this figure, the global estimates are fed back to the local filters.*

framework for fusing multi-modal sensor data for tracking is the various modifications and extensions of the Kalman filter (Brown & Hwang, 1997; Kalman, 1960; Welch & Bishop, 2001). The Kalman filter is a recursive estimation method that tries to estimate the state of a discrete-time controlled process (i.e., the pose of a tracked object, possibly with velocity and acceleration information) using observable measurements (i.e., data from tracker sensors). The state is estimated in each time step by a set of update equations in the form of "predict" and "correct" cycle. The typical Kalman filter formulation requires that all the state variables for the available devices and the relevant equations be set up globally at the beginning. Given the dynamic and distributed nature of the framework described in this article, this approach is not practical. Instead, the federated Kalman filter first described by Carlson (1990) is to be used. In this framework, the sensors have their own local Kalman filter running that estimates the local state and covariance. Then for each object that is tracked, there is a master Kalman filter that uses the estimates coming from the local Kalman filters and computes a global estimate of the state and covariance. The results of the master filter can be fed back to the local filters to improve their local state estimates also. The rough architecture of such a federated Kalman filter is shown in Figure 2.

The federated Kalman filter allows for the assembling of a master filter depending on the new set of sensors found through the discovery process. The covariance matrix is a measure of the error in the state information and can be used as part of the QoS information.

FUTURE TRENDS

The future plans include testing the methods by an integrated system that consists of multiple trackers distributed over a sufficiently large area. A variety of trackers such as vision-based trackers (e.g., AR Toolkit), Wi-Fi based trackers, magnetic trackers, and inertial trackers will be utilized in testing this prototype system. Application prototypes will be created to show the effectiveness of the proposed research. These prototypes will act as the test-beds and will provide feedback to the principles of the proposed discovery service.

CONCLUSION

The software framework outlined in this article is a promising approach to developing practical wide-area tracking systems that can utilize existing tracker infrastructures. The limitations of the framework are mainly due to the hardware. For example, one cannot foresee all the possible trackers and thus equip the tracked object accordingly ahead of time. In order to accommodate this, the framework assumes multiple, heterogeneous *classes* of tracking systems rather than instances of trackers. An example of tracker class might be a vision-based tracker. Thus, there may be many instances of such trackers, but one needs to equip the object with a standard fiducial marker.

REFERENCES

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., & MacIntyre, B. (2001). Recent advances in augmented reality. *IEEE Computer Graphics & Applications, 21*(6), 34-47.

Bahl, P., & Padmanabhan, V. (2000). RADAR: An in-building RF-based user location and tracking system. *Proceedings of the IEEE Infocom 2000* (pp. 775-784). IEEE CS Press.

Balachandran, A., Voelker, G., & Bahl, P. (2003). Wireless hotspots: Current challenges and future directions. *Proceedings of the 1st ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots* (pp. 1-9). ACM Press.

Bauer, M., Bruegge, B., Klinker, G., MacWilliams, A., Reicher, T., Riss, S., Sandor, C., & Wagner, M. (2001) Design of a component-based augmented reality framework, In *Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR 2001)* (pp.45-54).

Bauer, M., Bruegge, B., Klinker, G., MacWilliams, A., Reicher, T., Sandor, C., & Wagner, M. (2002). An architecture concept for ubiquitous computing aware wearable computers. *Proceedings of the International Workshop on Smart Appliances and Wearable Computing.*

Beugnard, A., Jezequel, J., Plouzeau, N., & Watkins, D. (1999). Making components contract aware. *IEEE Computer, 32*(7), 38-45.

Borriello, G., Chalmers, M., LaMarca, A., & Nixon, P. (2005). Delivering real-world ubiquitous location systems. *Communications of the ACM, 48,* 36-41.

Breen, D., Whitaker, R., Rose, E., & Tuceryan, M. (1996). Interactive occlusion and automatic object placement for augmented reality. *Proceedings of the Eurographics '96 Conference* (pp. 11-22).

Brown, R., & Hwang, P. (1997). *Introduction to random signals and applied Kalman filtering* (3rd ed.). New York: John Wiley & Sons.

Carlson, N. A. (1990). Federated square root filter for decentralized parallel processes. *IEEE Transactions on Aerospace and Electronic Systems, 26*(3), 517-525.

Czarnecki, K., & Eisenecker, U. (2000). *Generative programming: Methods, tools, and applications.* Boston: Addison-Wesley.

Feiner, S., MacIntyre, B., Hoellerer, T., & Webster, T. (1997). A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Proceedings of the International Symposium on Wearable Computers (ISWC'97)* (pp. 74-81).

Foxlin, E., & Durlach. N. (1994). An inertial head-orientation tracker with automatic drift compensation for use with HMDs. *Proceedings of the VRST '94 Conference* (pp. 159-173), Singapore.

Harter, A., Hopper, A., Steggles, P., Ward, A., & Webster, P. (2002). The anatomy of a context-aware application. *Wireless Networks, 8*(2-3), 187-197.

Hightower, J., & Borriello, G. (2001). Location systems for ubiquitous computing. *Computer, 34*(8), 57-66.

Huang, Z. (2003). *The uniframe system-level generative programming framework.* Unpublished MS thesis, Department of Computer and Information Science, Indiana University Purdue University Indianapolis, USA. Retrieved May 11, 2006, from http://www.cs.iupui.edu/uniFrame/

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering,* 35-45.

Koller, D., Klinker, G., Rose, E., Breen, D., Whitaker, R., & Tuceryan, M. (1997). Real-time vision-based camera tracking for augmented reality applications. *Proceedings of the Symposium on Virtual Reality Software and Technology* (pp. 87-94).

Neumann, U., & Cho, Y. (1996). A self-tracking augmented reality system. *Proceedings of the ACM Symposium on Virtual Reality and Applications* (pp. 109-115).

Newman, J., Ingram, D., & Hopper, A. (2001). Augmented reality in a wide area sentient environment. *Proceedings of the International Symposium on Augmented Reality* (pp. 77-86), New York. IEEE Press.

Newman, J., Wagner, M., Bauer, M., MacWilliams, A., Pintaric, T., Beyer, D., et al. (2004). Ubiquitous tracking for augmented reality. *Proceedings of the 3rd IEEE and ACM International Symposium on Mixed and Augmented Reality* (pp. 192-201).

Olson, A., Raje, R., Bryant, B., Burt, C., & Auguston, M. (2005). UniFrame: A unified framework for developing service-oriented, component-based, distributed software systems. In Z. Stojanovic & A. Dahanayake (Eds.), *Service oriented software system engineering: Challenges and practices* (Chapter IV, pp. 68-87). Hershey, PA: Idea Group Inc.

Raje, R. (2000). UMM: Unified Meta-object Model for open distributed systems. *Proceedings of the 4th IEEE International Conference on Algorithms and Architecture for Parallel Processing* (pp. 454-465). IEEE Press.

Reitmayr, G., & Schmalstieg, D. (2001). An open software architecture for virtual reality interaction. *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (pp. 47-54). ACM Press.

Schilit, B.N., LaMarca, A., Borriello, G., Griswold, W.G., McDonald, D., Lazowska, E., et al. (2003). Challenge: Ubiquitous location-aware computing and the "Place Lab" initiative. *Proceedings of the 1st ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots* (pp. 29-35). ACM Press.

Siram, N. (2002). *An architecture for discovery of heterogeneous software components.* Unpublished MS thesis, Department of Computer and Information Science, Indiana University Purdue University Indianapolis, USA. Retrieved May 11, 2006, from http://www.cs.iupui.edu/uniFrame/

State, A., Hirota, G., Chen, D., Garrett, W., & Livingston, M. (1996). Superior augmented reality registration by integrating landmark tracking and magnetic tracking. *Proceedings of SIGGRAPH '96* (pp. 429-438).

Tuceryan, M., Greer, D., Whitaker, R., Breen, D., Crampton, C., Rose, E., & Ahlers, K. (1995). Calibration requirements and procedures for a monitor-based augmented reality system. *IEEE Transactions on Visualization and Computer Graphics, 1*(3), 255-273.

Tuceryan, M., Genc, Y., & Navab, N. (2002). Single Point Active Alignment Method (SPAAM) for optical see-through (HMD) calibration for augmented reality. *Presence: Teleoperators and Virtual Environments, 11*(3), 259-276.

Ubisense. (2006). Retrieved May 3, 2006, from http://www.ubisense.net/

Welch, G., & Bishop, G. (2001). An introduction to the Kalman filter. *Proceedings of the Siggraph Course,* Los Angeles.

Welch, G., & Foxlin, E. (2002). Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications, 22*(6), 24-38.

Whitaker, R., Crampton, C., Breen, D., Tuceryan, M., & Rose, E. (1995). Object calibration for augmented reality. *Proceedings of Eurographics'95* (pp. 15-27).

KEY TERMS

Augmented Reality: Superimposing information on the view of the physical world for the purposes of providing information.

Degrees-of-Freedom (DOF): For a tracker of a particular type, the number of independent dimensions of information obtained from the sensor hardware.

Kalman Filter: A linear estimation technique first proposed by Rudolph Kalman in 1960 that is extensively used in tracking and navigation applications.

Tracking: A hardware/software system that can provide the position and/or orientation of an object being tracked in real time.

UniFrame: A unifying framework that supports a seamless integration of distributed and heterogeneous components.

UniFrame Resource Discovery System (URDS): A system that provides an infrastructure for proactively discovering components deployed over a network.

Distributed Web GIS

Jihong Guan
Tongji University, China

Shuigeng Zhou
Fudan University, China

Jiaogen Zhou
Wuhan University, China

Fubao Zhu
Wuhan University, China

INTRODUCTION

The popularity of World Wide Web and the diversity of GISs on the Internet have led to an increasing number of geo-referenced information (GRI) sources that spread over the Internet. How to integrate the heterogeneous and autonomous GISs to facilitate GRI accessing, data sharing, and interoperability is still a big challenge. Furthermore, the rapidly emerging mobile Internet and constantly increasing number of wireless subscribers bring new opportunities to geographic information services. Putting the Internet GIS in the palm will enable us to access geographic information with personal devices anytime and anywhere.

In the past decade, a lot of research has been done on designing interoperable systems in which collections of autonomous and heterogeneous GISs can cooperate to carry out query tasks. However, as far as system architecture is concerned, current solutions for integration of distributed GIS applications are mainly based on either C/S or B/S mode. The inherent limitations of these modes—for example, requiring a proper bandwidth, high-quality and stable network connection, less supporting of group awareness, and high-level cooperation—make them incompetent to fulfill various requirements of a dynamic, complicated, and distributed network computing environment, especially the mobile network environment, where the wireless communication networks have low bandwidth, frequent disconnections, and long latency, and the mobile devices (PDAs or mobile phones) have limited power, memory, computational power, and displaying capability. Such a situation calls for a new framework to support globally geographic information accessing and sharing in the (mobile) Internet environment.

The mobile agent is a recently developed computing paradigm that offers a full-featured infrastructure for development and management of network-efficient applications. Mobile agents are processes dispatched from one host to another during its execution on behalf of its owner or creator to accomplish a specified task. Agent-based computing can benefit Internet (especially mobile Internet) applications by *providing asynchronous task execution and more dynamics, supporting flexible and extensible cooperation, reducing communication bandwidth, enhancing real-time abilities and a higher degree of robustness, enabling off-line processing and disconnected operation.* Thus it is natural to introduce mobile agents into accessing and sharing distributed geographic information in a (mobile) Internet environment.

This article presents the MADGIS (Mobile Agent-based Distributed Geographic Information System) project, which aims at integrating distributed Web GIS applications by using mobile agent technologies to overcome the limitations of traditional distributed computing paradigms in a (mobile) Internet context.

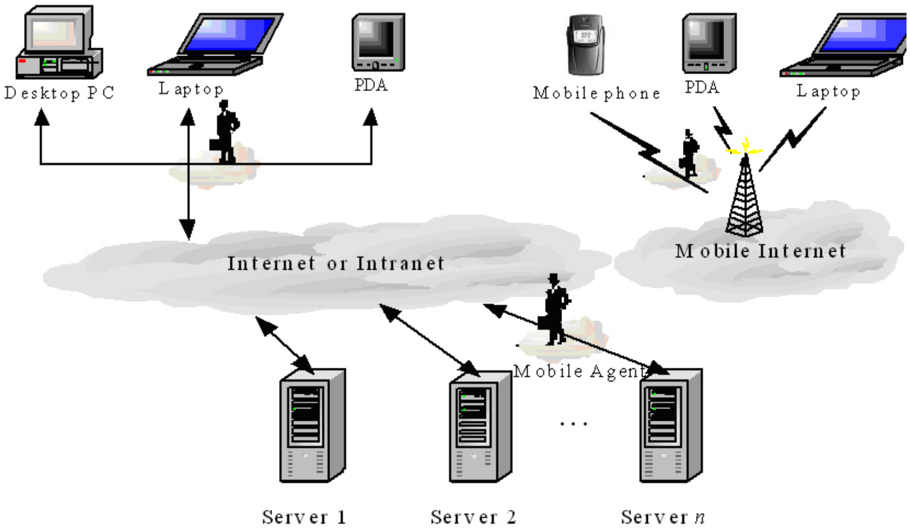
MADGIS FRAMEWORK

The MADGIS system consists of client sites (or clients), sever sites (or servers), a (mobile) Internet or intranet connecting these sites, and mobile agents roaming on the Internet/intranet for retrieving information on behalf of the clients. Figure 1 is an overview of MADGIS.

In MADGIS, a client site refers to a client machine, which can be a desktop, a laptop personal computer, a PDA, or a mobile phone used for query submission and results presentation. A server site is also a MADGIS server that provides spatial information services for local or remote requests. A user submits a query from a client machine to a server *via* Web browser. The query is analyzed and optimized by the server, from which one or multiple mobile agents are created and dispatched to accomplish the query task cooperatively. Each mobile agent along with its sub-task travels from one remote server to another to gather the related information. Retrieved information is then taken back to the original site after the mobile agent finishes its mission. All returned

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. MADGIS overview



information is further merged there and presented to the user. The servers also provide a docking facility for mobile agents in case they cannot travel back to the destinations promptly due to network problems.

The Client Site of MADGIS

A user can access any GISs within the MADGIS system via a client or a local server. First, a user should log into one server in the system. Then the server returns a Web (HTML) page to the client, in which there is a Java Applet termed as Client-Applet composed of one mobile agent environment (MAE), one stationary agent, and one mobile agent. The client-applet is executed at the client site to establish the MAE for the client and to start the stationary agent encoded in the Client-Applet. We call this stationary agent "client-agent", which is responsible for two tasks:

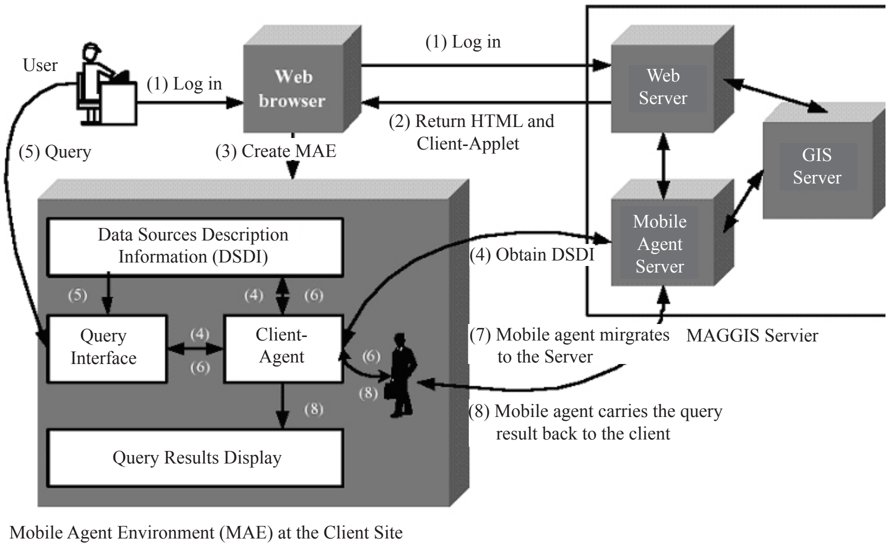
1. To obtain the data sources description information (DSDI) from the visited server, which includes all data sources' metadata (e.g., names, URLs, and schema of each data source). Users submit queries to or browse the MADGIS system according to the DSDI. The client-agent gets DSDI from the stationary agent of the visited server. At each server, DSDI is maintained by the local stationary agent. Besides responsibility for maintaining local DSDI, the stationary agent should also send messages to other servers in the system to notify of updates of DSDI, so as to keep the global DSDI updating simultaneously.
2. To create the query interface (QI) in the Web browser with which the user submits queries and gets retrieved data.

Thus when the query environment is set up at the client site, what a user can see is only the QI, while client-agent, mobile agent, and its execution environment are at the back-end. The user starts his or her query operations via QI, and the server accessed will take charge of query processing and mobile agent manipulation. Typically, a whole query session consists of the following steps:

1. At a client site, a user visits one server *via* Web browser by specifying the server's URL.
2. The accessed server returns a Web page including a client-applet.
3. The client-applet is executed at the client to establish MAE and to start the stationary client-agent.
4. The client-agent obtains the DSDI from the server and creates the QI for the user.
5. The user constructs his or her query and submits it via the QI to the server.
6. When the client-agent gets the user's query, it initiates a mobile agent to take over the query task.
7. The mobile agent with user's query task migrates to the server to which the client first visited for further query processing.
8. After the query task is completed at the server, the mobile agent moves back to the client and returns the results to the user via the browser.

Above, steps 1 to 4 are necessary for a client to access MADGIS. After that, the client can submit queries that are answered by following steps 5 to 8 repeatedly. The process described above and the interaction among client, server, and mobile agent are demonstrated in Figure 2.

Figure 2. The client site of MADGIS


User
(1) Log in
Web browser
(1) Log in
(2) Return HTML and Client-Applet
Web Server
GIS Server
Mobile Agent Server
(5) Query
(3) Create MAE
Data Sources Description Information (DSDI)
(4) Obtain DSDI
(5)
(4)
(6)
Query Interface
(4)
(6)
Client-Agent
(6)
(8)
(7) Mobile agent mirgrates to the Server
MAGGIS Servier
(8)
(8) Mobile agent carries the query result back to the client
Query Results Display
Mobile Agent Environment (MAE) at the Client Site

The Server Site of MADGIS

A server in MADGIS consists of at least three main components: Web server (or simply WSever), GIS server (GSever), and mobile agent server (MAServer). WServer is the interface for a user to connect to a MADGIS server and responsible for providing client-applet to the user. GServer is composed of a GIS database and a spatial query processing engine that provides support for local query processing. MAServer is the key component of a MADGIS server; it provides the facilities to support mobile agents to carry out query processing. Figure 3 illustrates the architecture of a typical MADGIS server. In what follows, we discuss MAServer in detail.

A MAServer contains four stationary agents: local services agent (LSA), query optimization agent (QOA), querying and wrapping agent (QWA), and mediation and transformation agent (MTA).

LSA has the following duties: (a) providing the system's DSDI for clients who log into the server; (b) maintaining local DSDI and notifying local DSDI updates to remote servers; (c) maintaining the statistical information of local data resources; and (d) sensing and detecting the status of network traffic between the local server and remote servers or clients.

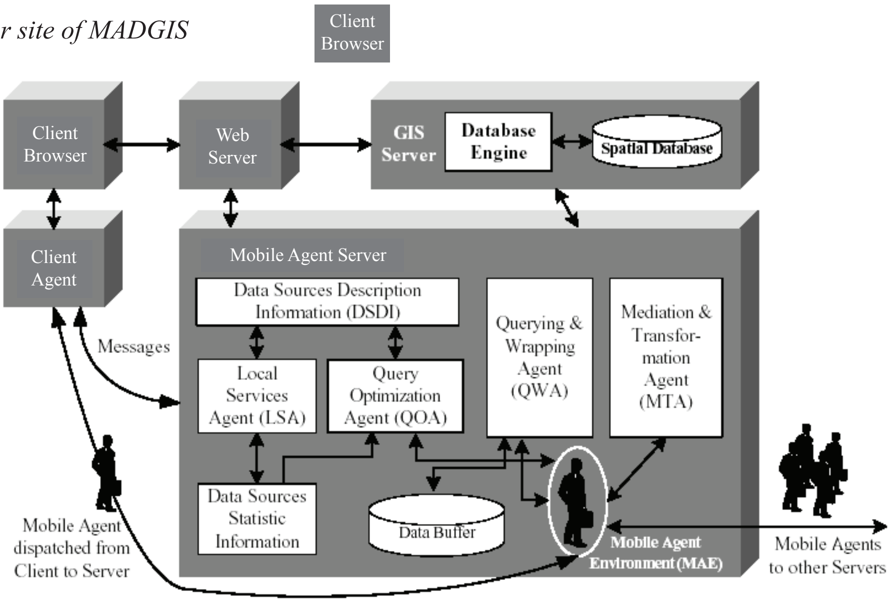
QOA is responsible for analyzing and optimizing the user's query. A user's query first performs grammar-check and is parsed by the Client-Agent; it is then taken to the server site by a mobile agent (termed main query agent). The QOA at the server site takes charge of the query's optimization and determines query strategy accordingly. A user query may be split into several sub-queries, each of which is related to one server's data source.

QWA is responsible for fulfilling the sub-query task related to local data and wrapping the query results into a standard GML document. When a user query involves data of other data sources or sites, data retrieved from multiple sites has to be merged, which is done by MTA. Then the merged query results will be further transformed into a SVG document and taken back to the client by mobile agent.

Figure 3 also illustrates the procedure of how a query task is completed by using mobile agent. The procedure includes the following steps:

1. The query task is shifted to QOA for optimization to reduce computational cost and network transmission volume incurred by query processing. After optimization, a query plan is created, which includes a set of sub-queries, sites on which the sub-queries are executed respectively, and the execution sequences of these sub-queries. QOA returns the result of query optimization to the main query agent. The main query agent then decides whether additional mobile agents are requested to carry out sub-queries processing.
2. If the query involves only local data, the main query agent will go on to finish the query task itself without necessity of spawning other mobile agents. The main query agent assigns the query task to QWA, who is in charge of retrieving data and wrapping the results.
3. If the query involves data of multiple server sites, and the sub-queries are requested to evaluate in sequence, then the main query agent or a mobile agent created by the main query agent will take over the query task. The mobile agent will be dispatched out according to its itinerary arranged previously. Since each sub-query involves one remote site, the sub-queries will

Figure 3. The server site of MADGIS


Client Browser
Client Browser
Web Server
GIS Server
Database Engine
Spatial Database
Client Agent
Mobile Agent Server
Data Sources Description Information (DSDI)
Querying & Wrapping Agent (QWA)
Mediation & Transformation Agent (MTA)
Messages
Local Services Agent (LSA)
Query Optimization Agent (QOA)
Data Sources Statistic Information
Data Buffer
Mobile Agent dispatched from Client to Server
Mobile Agent Environment (MAE)
Mobile Agents to other Servers

be accomplished one by one. After the last sub-query is finished, the dispatched mobile agent will go back to the client site directly or to its original server site.

4. Otherwise, if the query involves data of multiple sites, and the sub-queries are requested to evaluate in parallel, and the local site and other remote sites corresponding to the sub-queries are connected, then multiple mobile agents will be cloned by the main query agent to execute the sub-queries in parallel so as to gain better efficiency. The main query agent may join the mobile agents group to finish the query task or just take the role of coordinator of the multiple agents.
5. Query results obtained from all related sites are brought back to the original server, then MTA at the server site will do data integration and transformation, and one SVG document as the final result of the query will be created.
6. The main query agent carries the final query result in SVG format to the client, which is presented to the user via browser.

Another function of the MADGIS server is to provide the docking mechanism for mobile agents when connection between the current site and the destination site of agent migration is disrupted. LSA will take the role of deactivation and activation of mobile agents when such a situation happens.

Mobile Agent Environment

The mobile agent environment (MAE) exists at both the client and the server sites. It provides an environment for mobile agents to create, execute, dispatch, and migrate. Besides the mobile agents, MAE is composed of the following functional modules: mobile agent manager (MAM), mobile agent transportation (MAT), mobile agent naming (MAN), mobile agent communication (MAC), and mobile agent security (MAS).

MAM, the heart of MAE, is responsible for all kinds of management of mobile agents. It mainly provides a full-fledged environment for agent creating and executing, basic functions to make the mobile agent migrate precisely to its destination, functions for agent scheduling locally, and support for an agent's remote management. MAT controls the transferring of mobile agents—that is, sending and receiving mobile agents to and from remote sites. MAN manages a mobile agents' naming service, which provides the mechanism of tracing mobile agents. MAC serves mobile agent communication, which serves as the protocol of communication, collaboration, and events transmission among agents. MAS provides a two-facet security mechanism. On the one hand, it is responsible for distinguishing users and authenticating their mobile agents in order to protect server resources from being illegally accessed or even maliciously attacked. On the other hand, it ensures mobile agents not be tampered by malicious hosts or other agents.

MADGIS PROTOTYPE

We use Aglet Software Development Kit (ASDK) 2.0 (http://www.tr1.ibm.co.jp/aglets/), JDK 1.3, and Geotools 0.8.0 for implementing the MADGIS prototype. Geographic information of states, cities, rivers, roads, and lakes of the United States, Canada, and Mexico from ArcView GIS 3.2 samples and Guangzhou (a city in China) Agricultural Information Systems are processed and stored on five computers, which take the roles of servers. Users can access MADGIS transpar-

Figure 4. Query result 1 of the prototype

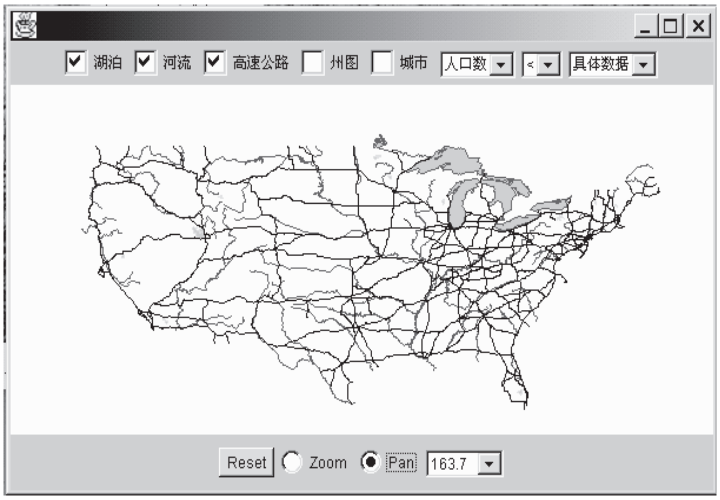
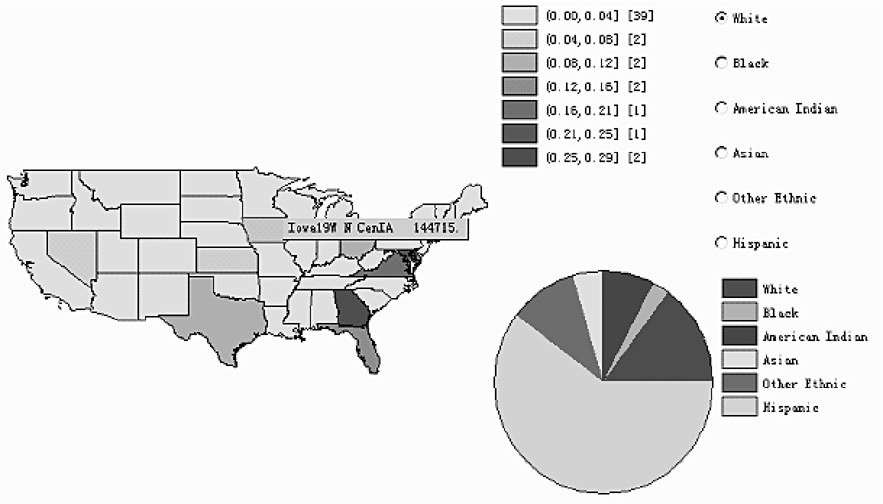

☑ 湖泊
☑ 河流
☑ 高速公路
☐ 州图
☐ 城市
人口数
<
具体数据
Reset
○ Zoom
● Pan
163.7

Figure 5. Query result 2 of the prototype


[0.00,0.04] [39]
[0.04,0.08] [2]
[0.08,0.12] [2]
[0.12,0.16] [2]
[0.16,0.21] [1]
[0.21,0.25] [1]
[0.25,0.29] [2]
● White
○ Black
○ American Indian
○ Asian
○ Other Ethnic
○ Hispanic
Iowa19W N CenIA 144715.
White
Black
American Indian
Asian
Other Ethnic
Hispanic

ently from any client machine connected to the Internet for geographic information stored in the system. Figures 4 and 5 show the interface of the prototype and the query results for roads, rivers, and lakes, and population distribution.

RELATED WORK

We give a brief survey on related work of distributed GISs and mobile agent areas in this part.

For distributed GISs, quite a lot research has been done in the past decade to design distributed interoperable systems (Leclercq, Benslimane, & Yetongnon, 1997; Fonseca & Egenhofer, 1999; Abel, Ooi, Tan, & Tan, 1998; Wang, 2000), which can be classified into three levels of integration: platform level, syntactical level, and application level (Fonseca & Egenhofer, 1999). Platform-level integration is concerned with hardware, operating systems, and network protocols, providing support for transferring of flat structure files between systems. Syntactic-level integration provides functionalities and tools for defining persistent and uniform views over multiple heterogeneous spatial data sources. Application-level integration aims at defining seamless system integration. However, these existing approaches and systems are mainly based on a connection-oriented mechanism (C/S and B/S modes) not suitable for efficient accessing, sharing, and integration of distributed systems in a (mobile) Internet environment.

For mobile agent-based applications, as a new and promising distributed computing paradigm, mobile agents have been employed in many applications, which include information retrieval (Brewington et al., 1999), Web search (Kato, 1999), electronic commerce (Dasgupta, 1999), and personalization Web services (Samaras & Panayioyou, 2002). Tsou and Buttenfield (2000) proposed to use agents to provide distributed GIS services; however, agents referred to in this article are *static* agents, which are quite different from the *mobile* agents employed in this article. A static agent is just a program with a certain autonomy or even intelligence, but it cannot migrate from one site to another to carry out missions on behalf of its own.

CONCLUSION

The Internet has greatly changed the ways of geographic data accessing, sharing, and disseminating. The emergence of mobile agent technologies brings new opportunities and challenges to Web-based geographic information services. To fulfill the requirements of distributed GIS accessing and sharing under a (mobile) Internet environment, we propose the MADGIS framework for accessing distributed Web-based GISs by using mobile agent technologies. By introducing mobile agents to carry out query tasks, MADGIS can save significant bandwidth by moving locally to the resources needed; carry the code to retrieve remote information, without needing the remote availability of a specific server; proceed without continuous network connections, for interacting entities can be moved to the same site when connections are available, and can then interact without requiring further network connections; and work with mobile computing systems (e.g., laptop, palmtop).

ACKNOWLEDGMENTS

This work was supported by grants numbered 60573183 and 60373019 from NSFC, grant No. 20045006071-16 from the Chenguang Program of Wuhan Municipality, grant No. WKL(04)0303 from the Open Researches Fund Program of LIESMARS, and the Shuguang Scholar Program of Shanghai Education Development Foundation.

REFERENCES

Abel, D., Ooi, B., Tan, K., & Tan, S.H. (1998). Towards integrated geographical information geoprocessing. *International Journal of Geographical Information Science,* 353-371.

Brewington, B., Gray, R., Moizumi, K., Kotz, D., Cybenko, G., & Rus, D. (1999). Mobile agents in distributed information retrieval. In M. Klusch (Ed.), *Intelligent Information Agents* (pp. 355-395). Berlin: Springer-Verlag.

Dasgupta, P. (1999). MagNET: Mobile agents for networked electronic trading. *IEEE Transactions on Knowledge and Data Engineering, 11*(4), 509-525.

Fonseca, F., & Egenhofer, M. (1999). Ontology-driven geographic information systems. *Proceedings of the 7th ACM Symposium on Advances in Geographic Information Systems* (pp. 14-19), Kansas City.

Geo-Agent. (n.d.). Retrieved from http://map.sdsu.edu/geo-agent/

Kato, K. (1999). An approach to mobile software robots for the WWW. *IEEE Transactions on Knowledge and Data Engineering, 11*(4), 526-548.

Leclercq, E., Benslimane, D., & Yetongnon, K. (1997). Amun: An object-oriented model for cooperative spatial information systems. *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop* (pp. 73-80), Newport Beach, CA.

Open GIS Consortium (OGC). (n.d.). Retrieved from http://www.opengis.org/

Sahai, A. (n.d.). *Intelligent agents for a mobile network manager (MNM).* Retrieved from http://www.irisa.fr/solidor/doc/ps97/

Samaras, G., & Panayioyou, C. (2002). A flexible personalization architecture for wireless Internet based on mobile agents. *Proceedings of ADBIS 2002* (LNCS 2435, pp. 120-134).

Tsou, M.H., & Buttenfield, B.P. (2000). Agent-based mechanisms for distributing geographic information services on the Internet. *Proceedings of the 1st International Conference on Geographic Information Science,* Savannah, GA.

Wang, F. (2000). A distributed geographic information system on the Common Object Request Broker Architecture (CORBA). *Geoinformatica, 4*(1), 89-115.

World Wide Web Consortium. (1999, August). *Scalable Vector Graphics (SVG) 1.0 specification* (W3C work draft). Retrieved from http://www.w3.org/tr/svg

XSLT. (n.d.). Retrieved from http://www.w3.org/XSLT

KEY TERMS

Agent: Takes the roles of a representative and acts on behalf of other persons or organizations; is often a software routine that waits in the background and performs an action when a specified event occurs.

Client: A computer or program that can download files for manipulation, run applications, or request application-based services from a file server.

Data Sharing: The ability to share the same data resource with multiple applications or users. Implies that the data are stored in one or more servers in the network and that there is some software locking mechanism that prevents the same set of data from being changed by two people at the same time.

Geographic Information System (GIS): A computer application used to store, view, and analyze geographical information, especially maps. Often called "mapping software," it links attributes and characteristics of an area to its geographic location.

MADGIS: Mobile Agent-based Geographic Information System.

Mobile Agent: A composition of computer software and data which is able to migrate (move) from one computer to another autonomously and continue its execution on the destination computer with the features of autonomy, social ability, learning, and most important, mobility.

Mobile Agent Environment (MAE): Provides an environment for mobile agents to create, execute, dispatch, and migrate.

Server: A computer that processes requests for HTML and other documents that are components of Web pages. The term "server" may refer to both the hardware and software (the entire computer system) or just the software that performs the service.

Web GIS: Short for Web-based GIS; a geographic information system that deals with spatial information and provides it Web users via Web browsers.

Dynamic Pricing Based on Net Cost for Mobile Content Services

Nopparat Srikhutkhao
Kasetsart University, Thailand

Sukumal Kitisin
Kasetsart University, Thailand

INTRODUCTION

In the past few years, the mobile phone's performance has increased rapidly. According to IDC's Worldwide Mobile Phone 2004-2008 Forecast and Analysis, sales of 2.5G mobile phones will drive market growth for the next several years, with sales of 3G mobile phones finally surpassing the 100 million annual unit mark in 2007. Future mobile phones can support more than 20,000 colors. With the advancements in functionality and performance of mobile phones, users will use them for all sorts of activities, and that will increase mobile content service requests. Currently, the pricing of mobile content service is up to each provider; typically they implement a fixed price called a market price because the providers do not have a formula to estimate the price according to the actual cost of their services. This article proposes a dynamic pricing model based on net cost for mobile content services.

BACKGROUND

A mobile phone today can support various format data causing mobile content service popularity among all mobile phone users. They can request a music VDO clip, a song, or a mobile phone game program. The price of each mobile content service differs for each different format of data. For example, the price of a true-tone ring tone is 35 baht (Sanook.com, 2005), while a Java game download costs 40 baht (Siam2you, 2005).

Conventionally, an operator set a fixed market price for each mobile content service. The prices can vary from operator to operator. The pricing has not been calculated based on the net cost for the requested service. Therefore, the set price can be lower or much higher than the actual cost. To come up with a way for a provider to be able to set a mobile content service price based on its actual cost, the provider must be able to quantify its actual cost for service. This article presents mobile content service interaction models and formulas for estimating the actual cost of a mobile content service; a provider can refer to these models and formulas when pricing its services.

Data Formats

The previous section discusses improving the performance of mobile devices and the variety of content available. We can classify mobile content service into four types: audio, image, video, and application (ClearSky Mobile Media, 2005). Users can request an audio clip and use it as their ring tone. They can leave voice messages for each other or download an mp3 song for their entertainment (Nokia, 2005; Sony Ericsson, 2005; Samsung, 2005). The audio content can be of three sub-types: monophonic (Sonic Spot, 2005), polyphonic (Cakewalk, 2005), and true tone. Image format can be either static or dynamic/animation. Users can request a music VDO clip and play it on their mobile phones, and apparently, a few companies have started to provide NetTV on mobile phones as well. Lastly, an example of application content users widely request could be a Java game application.

Parties in Mobile Content Services

Providing mobile content services involves many parties. We consider the following five participants (Bratsberg & Wasenden, 2004; Andreas, 2001). The first party is a user requesting mobile content services. The second party is a mobile operator (MO), which is the owner of a mobile phone service frequency. When a user requests mobile content service, the user will send a request to his or her content provider through the MO's network. The third party is a content provider (CP), which is an organization to serve mobile contents. The MO may or may not have license on the contents. The fourth party is the content owner (CO), which could be a person or an organization that has authorization for legal distribution of the mobile contents. And the last party is a content aggregator (CA), a middleman between a user and a content provider. The CA can help increase the channel to serve mobile content services.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Mobile content service model 1

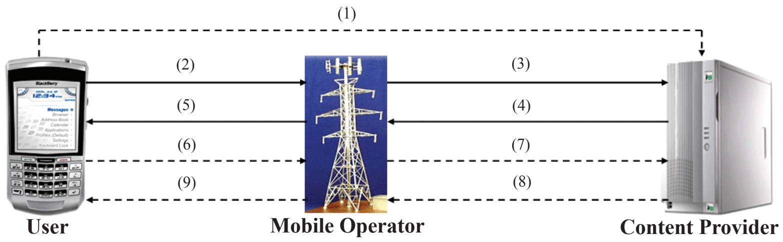
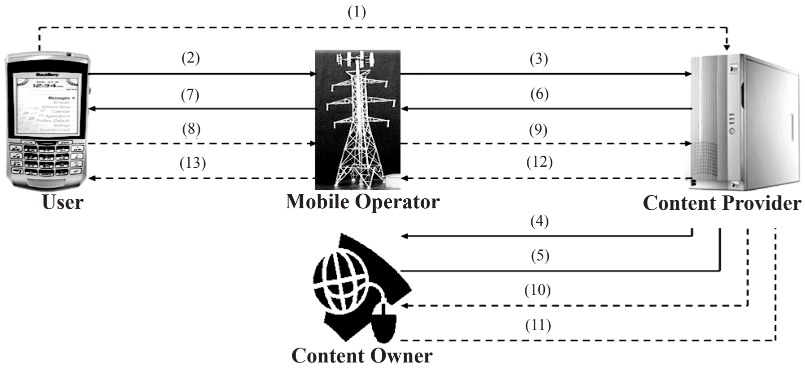


Figure 2. Mobile content service model 2



Request and Response Formats

When a user wants to use a mobile content service, he or she makes a request for the desired content from a CA or CP. Then the CA or CP responds to the user with the requested content. Requests and responses can be of the following four types: a Web request through a Web page, a short message service (SMS), an interactive voice responder (IVR), or a WAP request via a mobile internet WAP page. When the CA or CP responds successfully, the response can be sent using one of these four formats: a short message service (SMS), a smart message, a WAP push format, and a multimedia message service (MMS).

Mobile Content Service Models

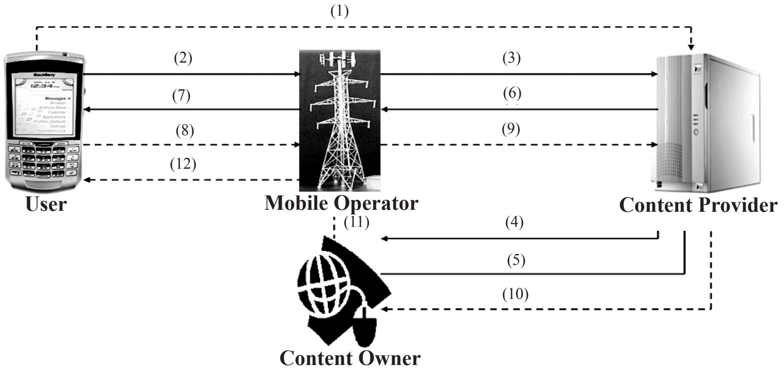
We categorized all mobile content services into the following interaction models based on involved parties and content providing methods.

Model 1: Parties involved are user, MO, and CP. A user requests content from a CP by using an SMS, IVR, WAP, or Web request. For any request format except a Web request, the request is sent to the CP through the MO's network. For a Web request format, the request is transferred to the CP directly. After the CP processes the request, the CP will reply to the user with the requested content information in the form of a WAP push, a WAP URL, or a bookmark. For a monophonic ring tone request, the CP will send content information in a smart message format. Other content formats can be replied to with an SMS, a WAP push, or an MMS. The workflow of model 1, as shown in Figure 1, is:

1. User requests content via a Web page.
2. User requests content via SMS, IVR, or WAP page.
3. MO forwards the request to CP.
4. CP sends the content information to user through MO.
5. MO forwards the content information from CP to user.
6. User connects to WAP page or open WAP push for retrieving the content file through MO.
7. MO redirects the file request to CP.
8. CP sends the content file to the user through network of MO.
9. MO transfers content file to the user.

Model 2: Parties involved are user, MO, CP, and CO, with the CP as the content file sender. A user sends a request to a CP in SMS, IVR, WAP or Web format. The CP, after receiving the request, sends this request to a CO for content information. The CO sends content information to the user via the CP in smart message format, SMS (URL for retrieving

Figure 3. Mobile content service model 3



content file), or a WAP push. After that, the CP forwards the information to user. The workflow of model 2, as shown in Figure 2, is as follows. Steps 1-3 are the same as in model 1. In step 4, the CP forwards the request to the CO. In step 5, the CO sends the content information to the user via the CP. Steps 6-9 of this model are the same as steps 4-7 of model 1. In step 10, the CP forwards the request to the CO. In step 11, the CO sends the content file to the user via the CP. In step 12, the CP sends the content file to the user through the network of the MO. In step 13, the MO transfers the content file to the user.

Model 3: Parties involved are user, MO, CP, and CO, with the CO as the content file sender. The CO has permission to distribute the content files. The CP is a middleman between the CO and the user. The user requests content from the CP, which then sends the request to the CO. The CO processes the request and sends the content file directly to the user. The workflow for model 3 is: steps 1-10 are the same as steps 1-10 of model 2. In step 11, the CO sends the content file to the user through the network of the MO. In step 12, the MO transfers the content file to the user.

Model 4: Parties are user, MO, CA, CP, and CO, with the CA as the content file sender. For model 4, there is a middleman between the user and the CP. When the user requests a service, the user sends a request to the CA. Then the request is forwarded to the CP and the CO. After the CO processes the request, the content file will be sent to the user via the CP and the CA. The workflow of model 4 is as follows:

1. User requests content via a Web page.
2. User requests content via SMS, IVR, or WAP page.
3. MO forwards request to CA.
4. CA forwards request to CP.
5. CP forwards request to CO.

6-9. CO sends content information to user via CP, CA, and MO.

10-13. User connects to WAP page or open WAP push for retrieving the content file through MO, CA, and CP.

14-17. CO sends the content file to user via CP, CA, and MO.

Model 5: Parties involved are user, MO, CA, CP, and CO, with the CO as the content file sender. Methods for requesting content in this model are the same as those of model 4. They can be via an SMS, IVR, or WAP request. Requests will be sent via the network of the MO. Another

Figure 4. Mobile content service model 4

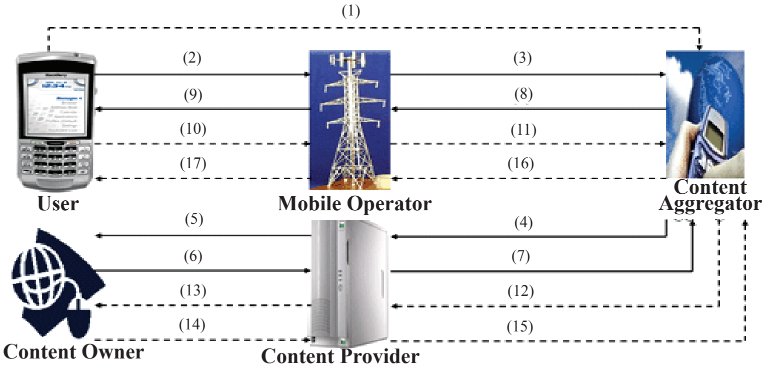
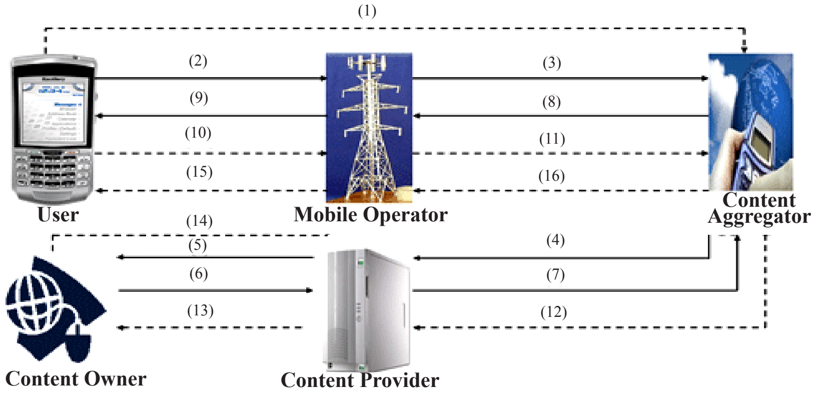


Figure 5. Mobile content service model 5



option is that a request can be sent to the CA directly. When the CO finishes processing the request, the CP then sends the content file directly to the user. The workflow of model 5 is as follows. Step 1-13 are the same as those in model 4. In step 14, the CO sends the content file to the user through the network of the MO. In step 15, the MO transfers the content file to the user.

Cost of Mobile Content Service

To be able to calculate the actual cost of a service, the providers must know the actual cost of providing the service. Two factors contributing to the cost of providing a mobile content service are the cost of software or content for the content service and the operational cost. Each party has a different operational cost and pays a different content fee or has a different revenue sharing model for a mobile content service (Smorodinsky, 2002; Kivisaari & Luukkainen, 2003; Stiller, Reichl, & Leinen, 2001). Operating costs for the MO are mainly the bandwidth cost for sending the content to the user. For the CP, the operating costs can be the cost of the bandwidth, the operation cost, the revenue sharing or fee for the MO, and the revenue sharing or fee for the CO. For the CO, its operating costs are from the cost of the bandwidth and the operation cost. And the CA's operating costs come from the cost of the bandwidth, the operation cost, the revenue sharing or fee for the MO, and the revenue sharing or fee for the CP.

Formula for Calculating an Actual Cost

Formulas depend on the mobile content service interaction models, the format of the mobile content, and its transfer venues. Parameters used in the formulas are as follows:

- S refers to the software value or value of a content file.
- I_A refers to the operation cost of CA.
- I_P refers to the operation cost of CP.
- I_O refers to the operation cost of CO.
- I refers to the total operation cost.
- B refers to content file size (Bit).
- D_M refers to bandwidth cost per bit for MO.
- D_A refers to bandwidth cost per bit for CA.
- D_p refers to bandwidth cost per bit of CP.
- D_O refers to bandwidth cost per bit of CO.
- A refers to bandwidth cost for CA.
- P refers to bandwidth cost of CP.
- O refers to bandwidth cost of CO.
- M refers to bandwidth cost of MO.
- C refers to the sum of bandwidth cost for CP and MO.
- R_1 refers to revenue sharing for MO.
- R_{2_1} refers to content fee for CO.
- R_{2_2} refers to revenue sharing for CO.
- R_{3_1} refers to fee for CP.
- R_{3_2} refers to revenue sharing for CP.
- W refers to the sum of dynamic costs before calculation revenue sharing and content fee for CO.
- E refers to the sum of dynamic cost before calculation revenue sharing and content fee for CP.
- T refers to dynamic cost before calculate revenue sharing for MO.
- N refers to the actual cost.

Formula 1: for model 1:

$I_P = I$, $B^* D_p = P$, $B^* D_M = M$, $P+M = C$, $S+I+C = T$, $T/(1-R_1) = N$

Formula 2: for model 2:

$I_P + I_O = I$, $B^* D_O = O$, $B^* D_p = P$, $B^* D_M = M$, $(O+P+M) = C$,

$W + R_{2\_1} = T$ *OR* $W / (1 – R_{2\_2}) = T$, $T/(1- R_1) = N$

Formula 3: for model 3:
$I_P + I_O = I$, $B^* D_O = O$, $B^* D_M = M$, $O+M = C$, $S+I+C = W$,

$W + R_{2\_1} = T$ *OR* $W / (1 – R_{2\_2}) = T$, $T/(1- R_1) = N$

Formula 4: for model 4:

$I_A + I_P + I_O = I$, $B^* D_O = O$, $B^* D_p = P$, $B^* D_A = A$, $B^* D_M = M$, $(O+P+A+M) = C$

$S+I+C = W$, $W + R_{2\_1} = E$ *OR* $W /(1- R_{2\_2}) = E$, $E + R_{3\_1} = T$ *OR* $E / (1- R_{3\_2}) = T$

$T / (1– R_1) = N$

Formula 5: for model 5:

$I_A + I_P + I_O = I$, $B^* D_O = O$, $B^* D_M = M$, $O+M = C$, $S+I+C = W$,

$W + R_{2\_1} = E$ *OR* $W /(1- R_{2\_2}) = E$, $E + R_{3\_1} = T$ *OR* $E / (1- R_{3\_2}) = T$

$T / (1– R_1) = N$

RESULTS AND ANALYSIS

We analyzed our actual cost formulas presented above by doing experiments based on three different types of transmitting channels: ADSL, leased line, and satellite. Cost for transmitting file content is determined by an average rate from ISPs in Thailand (True Internet, 2005; LOXINFO, 2005; Internet KSC, 2005; Ji-NET, 2005; INET, 2005). Each party uses the same sending channel. For the operation cost, we randomly selected a cost. The random method is Gaussians, with a base cost of 0.8262 and deviation of 0.14711. The software/content cost is a randomly selected value ranging from 0.1058 to 2.1160 baht. The sending channel of our results is satellite. Figures 6 and 7 show the average actual costs of true tone. The average operation costs for CA, CP, and CO is 0.829624 baht. Figures 8 and 9 shows the monophonic actual cost. And the average operation cost for CA, CP, and CO is 0.821213 baht.

Figure 6 shows average costs of true tone content for all service models and the market price. The average costs of models 1-5 are 1.66, 6.66, 6.64, 20.02, and 19.95 respectively. The maximum actual costs are less than market about 0.6 times. Thus, we found the market price for true tone content overpriced.

Figure 7 shows the probability of the customer being willing to pay the cost price of true tone content. For Figure 7, we found customers almost willing to pay the cost price of model 1 and customers willing to pay about 60% of the market price.

Figure 8 shows the average costs for a monophonic through satellite compared with the market price. From this figure, we found the market price is more than the actual

Figure 6. Average actual cost for a true tone content service

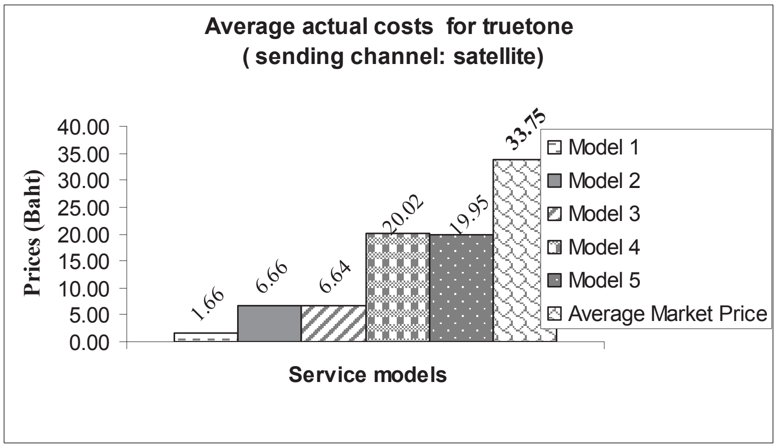


Figure 7. Probability of customer willing to pay the cost price of true tone content

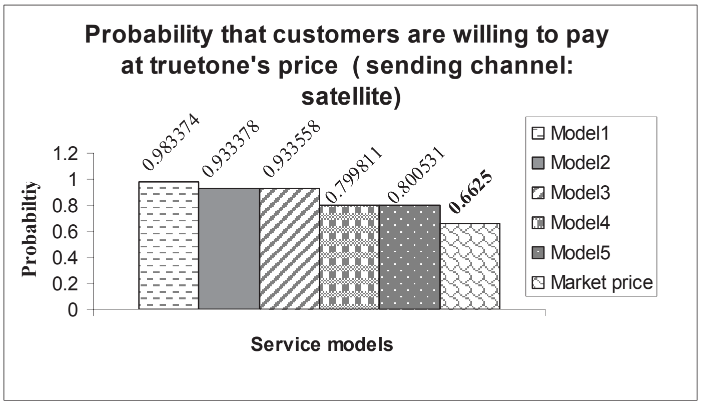


Figure 8. Average actual cost for a monophonic

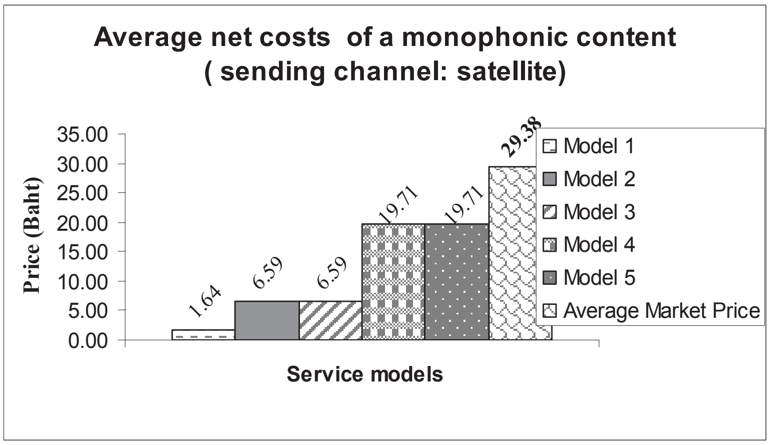
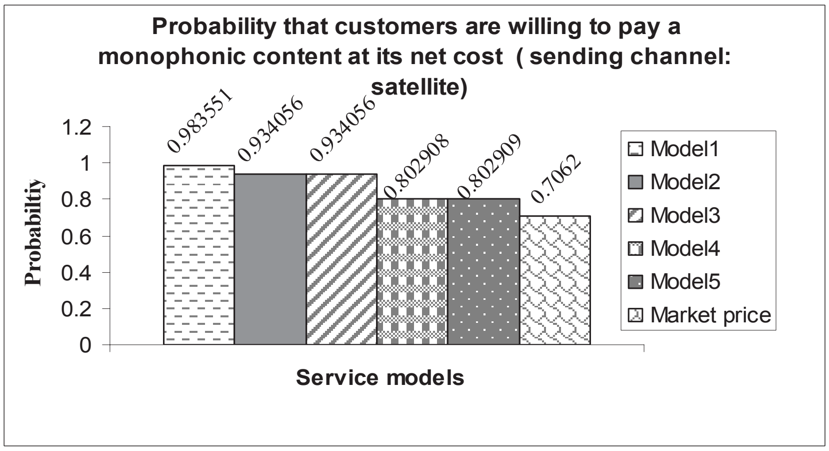


Figure 9. A probability that customers are willing to pay for monophonic content



costs for all models. Market prices are very expensive. It is about 18 times more than model 1's actual cost.

Figure 9 shows the probability that the customers are willing to pay for actual costs. We found customers will gladly pay the cost prices for model 1, model 2, and model 3. Also, we found the probability of customers willing to pay market price is less than the probability of model 1 by nearly 30%.

SUMMARY

From our preliminary results using our formulas, the size of file, number of parties, and the sender affect the calculation of the actual cost. We presented an alternative method in pricing the mobile content services based on the actual cost. The method allows a provider to dynamically set its service price accordingly.

REFERENCES

Andreas S. (2001). *The digital content network receiver service market.* White Paper, HTRC Group, Canada. Retrieved March 17, 2005, from http://www.htrcgroup.com/pdffiles/dcnr.pdf

Bratsberg, H., & Wasenden, O. (2004, September). *Changing regulation—Impacts on mobile content distribution.* Retrieved March 16, 2005, from http://web.si.umich.edu/tprc/papers/2004/373/bratsberg_wasenden_tprc04_mobile_content_distribution_final.pdf

Cakewalk. (2005). *Desktop music handbook: Glossary of MIDI and digital audio terms.* Retrieved August 28, 2005, from http://www.cakewalk.com/tips/desktop-glossary.asp

ClearSky Mobile Media. (2005). Retrieved July 13, 2005, from http://www.clearskymobilemedia.com/carriersol/en-cont.asp

INET. (2005). *Always by your side.* Retrieved August 25, 2005, from http://www.inet.co.th

Internet KSC. (2005). Retrieved August 25, 2005, from http://www.ksc.net

Ji-NET. (2005). Retrieved August 25, 2005, from http://www.ji-net.com

Kivisaari, E., & Luukkainen, S. (2003, March). Content-based pricing of services in the mobile Internet. *Proceedings of the 7th IASTED International Conference on Internet and Multimedia Systems and Applications.* Retrieved March 15, 2003, from http://www.tml.tkk.fi/~sakaril/Content-based_pricing.pdf

LOXINFO. (2005). Retrieved August 25, 2005, from http://www.csloxinfo.co.th/

Nokia. (2005). Retrieved August 20, 2005 from http://www.nokia.co.th/nokia/0,,51297,00.html

Samsung. (2005). *Digital world.* Retrieved August 23, 2005, from http://product.samsung.com/cgi-bin/nabc/product/b2c_product_type.jsp?eUser=&prod_path=/Phones+and+Fax+Machines%2fWireless+Phones

Sanook.com. (2005). Retrieved June 10, 2005, from http://mobilemagic.sanook.com

Siam2you. (2005). Retrieved June 10, 2005, from http://www.siam2you.com

Smorodinsky. R. (2002). *Mobile entertainment—A value chain analysis and reference business scenario.* Retrieved from http://www.fing.org/ref/upload/GlobalCommunicationsrevised.pdf

Sonic Spot. (2005). *Glossary.* Retrieved September 1, 2005, from http://www.sonicspot.com/guide/glossary.html

Sony Ericsson. (2005). *Products.* Retrieved August 18, 2005, from http://www.sonyericsson.com/spg.jsp?cc=global&lc=en&ver=4001&template=pg1_1&zone=pp

Stiller, B., Reichl, P., & Leinen, S. (2001, March). Pricing and cost recovery for Internet services: Practical review, classification and application of relevant models. *NETNOMICS: Economic Research and Electronic Networking, 3*(1). Retrieved January 12, 2005, from http://userver.ftw.at/~reichl/publications/NETNOMICS00.pdf

True Internet. (2005). Retrieved October 7, 2005, from http://www.asianet.co.th/home.htm

Efficient and Scalable Group Key Management in Wireless Networks

Yiling Wang
Monash University, Australia

Phu Dung Le
Monash University, Australia

INTRODUCTION

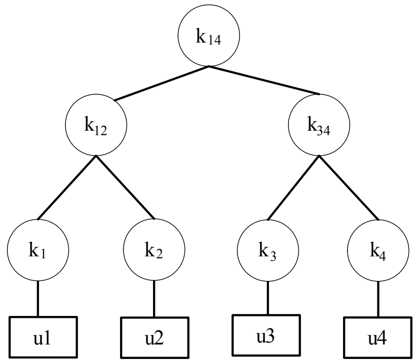
Multicast is an efficient paradigm to support group communications, as it reduces the traffic by simultaneously delivering a single stream of information to multiple receivers on a large scale. Along with widespread deployment of wireless networks and fast improving capabilities of mobile devices, it is reasonable to believe that the integration of wireless and multicast will result in enormous benefits. Before users can enjoy the flexibility and efficiency of wireless multicast, security must be achieved. The core issue of wireless multicast security is access control, which means that only authorized users can participate in the group communications. Access control can be achieved by encrypting the communication data with a cryptographic key, known as group key. Group key is shared by all the registered users, so that only authorized members can gain access to the group communication contents. Several group key management approaches (Challal & Seba, 2005; Sherman & McGrew, 2003; Kostas, Kiwior, Rajappan, & Dalal, 2003; Wong, Gouda, & Lam, 2000; Wallner, Harder, & Agee, 1999; Harney & Muckenhirn, 1997; Steiner, Tsudik, & Waidner, 1996; Mittra, 1997; Hardjono, Cain, & Monga, 2000; Kim, Perrig, & Tsudik, 2004; Perrig, Song, & Tygar, 2001) have been proposed in the literature, most of them directed towards wired networks. Although some approaches can be employed in the wireless environment, they cannot achieve the same efficiency as in the wired networks. The complexity of group key management in wireless networks cannot be confined only to the limitation of wireless networks such as higher data error rate and limited bandwidth, but also from the properties of wireless devices, such as insufficient computation power, limited power supply, and inadequate storage space.

BACKGROUND

Over the last decade, a large number of group key management approaches have been proposed. Among them, the most prominent is the logical key hierarchy (LKH) (Wong et al., 2000; Wallner et al., 1999). In LKH, a key tree is formed comprising group and other auxiliary keys (key encryption key, KEK) that are used to distribute the group key to the users. Figure 1 depicts a typical LKH key tree. In the LKH key tree, users are associated with the leaf nodes, and each user must store a set of keys along the path from leaf node up to the root. When membership changes such as join or departure, the rekeying procedure is invoked to update the keys along the path, thereby ensuring security. This update affects all the members in the tree. LKH algorithm has some drawbacks which prevent its application in the wireless environment:

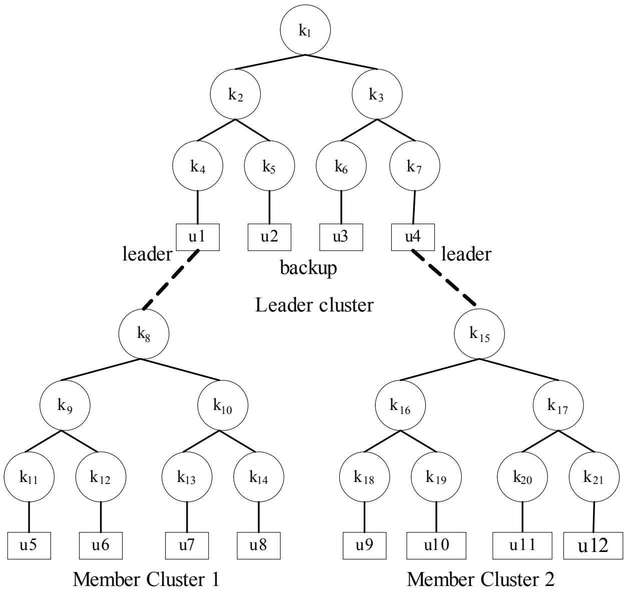
- **1-Affects-n:** As mentioned above, one membership change affects all the group members. But some changes are unnecessary to the members, especially in cellular networks, because the users in the cell are not only logical neighbors in the key tree but also physical (Sun, Trappe, & Liu, 2002).
- **Storage Inefficiency:** In the LKH algorithm, users have to store a set of keys. As the size of group increases, so does the number of keys stored by each user. This results in storage inefficiency of the lightweight mobile devices due to the limitation of storage space.

Figure1. Typical LKH key tree



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Group key management structure



Group Key Management Algorithm

Group key management algorithm is a core part of multicast security. It maintains the logical key structure and performs the procedures to assign, distribute, and update the group key and other KEKs.

Notation

In this section, we depict the notations that we will use in the following sections:

- bs: base station
- BS: a set of the base stations
- {x}k: message x is encrypted by the key k
- A → B {message}: A sends message to B via unicast
- A ⇒B {message}: A sends message to B via broadcast or multicast

The Proposed Logical Key Structure

In each cell of wireless network, the base station is responsible for managing the group key. In our proposed work the base station categorizes the authorized members into several clusters to form a two-level key management structure. Figure 2 depicts this logical structure within the cell. The group key management area is divided into smaller areas called clusters. Each cluster has its own cluster key for communication. The members of the leader cluster are assigned as leaders of the lower level member cluster—that is, one leader for one member cluster. The leader is responsible for distributing the rekeying messages to the lower level cluster users, thereby reducing the communication and computation overhead of the key server. The leftover users in the leader cluster serve as leadership backups.

The steps to build our proposed logical key management structure are as follows:

- **Step 1:** The base station groups users into several clusters based on the cluster policy, which defines the size of clusters and the ratio of leaders and backups. One cluster is assigned as the leader cluster, and others are member clusters.
- **Step 2:** Separate key trees are built for each cluster. The members in the leader cluster are assigned as leaders of member clusters—that is, one leader for one member cluster.
- **Step 3:** The base station assigns a local multicast address to each cluster for cluster communications.

The Proposed Group Key Management Algorithm

There are three main operations in the wireless group key management (multiple subgroups): member join, member leaving, and handoff. The rekeying procedures of these operations occur independently in each wireless cell. We illustrate our algorithm in each of these operations in the following subsections.

Member Join

When a user wants to join a group, backward secrecy must be maintained to prevent the new member from accessing the previous group communication details. The join procedure starts with the group join request sent by the user to the group key server (GKS).

u → GKS: {join request}

After authentication, GKS updates the group key and distributes to the base stations.

GKS ⇒ BS: {new group key}

There are two types of join: leader cluster join and member cluster join. The base station assigns the new member into a cluster according to the cluster policy, where leader cluster is given priority over member clusters. When the cluster is decided, the base station invokes the join procedure to rekey the cluster key tree. For example, in Figure 2, if u2 wants to join the group, then the rekeying procedure is invoked at the leader cluster. The base station needs to send the following two messages:

bs → u1: {new k_1, k_2, group key}k_4

bs ⇒ u3, u4: {new k_1, group key}k_3

u1 ⇒ cluster 1: {new group key}$k_{(\text{cluster-1})}$

u4 ⇒ cluster 2: {new group key}$k_{(\text{cluster-2})}$

The join procedure for the member cluster is similar to the leader cluster join. In Figure 2, if u6 wants to join the group, the rekeying is as follows:

bs → u5: { new k_8, k_9, cluster-1 key, group key}k_{11}

bs ⇒ u7, u8: {new k_8, cluster-1 key, group key}k_{10}

bs → u1: {new cluster-1 key, group key}k_4

bs ⇒ u2, u3, u4: {new group key}k_1

u4 ⇒ cluster 2: {new group key}$k_{(\text{cluster-2})}$

Member Leaving

When a member leaves the group, forward secrecy has to be guaranteed to keep the departing user from accessing the future group communication content. The leaving procedure can either be invoked by the user or initiated by the key server. In our proposal, there are two types of leaving: member cluster departure and leader cluster departure. First, we describe the user departure from member cluster, which is the most frequently occurring event. The rekeying procedure for this leaving is as follows:

GKS generates a new group key, and sends to base stations.

GKS ⇒ BS: {new group key}

The base station updates the affected keys in the member cluster key tree. For instance, in Figure 2, if u5 leaves the group, the base station sends four rekeying messages:

bs → u6: {new k_8, k_9, cluster-1 key, group key}k_{12}

bs ⇒ u7,u8: {new k_8, cluster-1 key, group key}k_{10}

bs → u1: {new cluster-1 key, group key}k_4

bs ⇒ u2, u3,u4: {new group key}k_1

u4 ⇒ cluster 2: {new group key}$k_{(\text{cluster-2})}$

As for the user departure from leader cluster, the situation becomes much more complex because of the need to select the next leader of the member cluster. This scenario can be classified into three categories:

1. The first scenario is when the leadership backup leaves the group. Since the backup has no association with the member cluster, the rekeying is limited to the leader cluster, followed by the rekeying procedure for the leaving operation. For example, in Figure 2, when u2 leaves the group:
   GKS generates a new group key and delivers it to the base stations.

   GKS ⇒ BS: {new group key}

   The base station updates the key tree of leader cluster.

   bs → u1: {new k_1, k_2, group key}k_4

   bs ⇒ u3, u4: {new k_1, group key}k_3

   u1 ⇒ cluster 1: {new group key}$k_{(\text{cluster -1})}$

   u4 ⇒ cluster 2: {new group key}$k_{(\text{cluster-2})}$

2. The second scenario is when the leader departs and the backup is to be elected as the next leader. Given a situation, the base station invokes the rekeying procedure in the leader cluster and assigns a backup to be the new leader of the affected cluster. For instance, as shown in the Figure 3, if u1 leaves the group, the base station assigns u3 to be the new leader of member cluster 1.

   bs → u2: {new k_1, k_2, group key}k_5

   bs ⇒ u3, u4: {new k_1, group key}k_3

   bs→ u3: {new cluster-1 key}k_6

   bs ⇒ cluster 1: {new cluster-1 key, group key}k_8

   u4 ⇒ cluster 2: {new group key}$k_{(\text{cluster-2})}$

3. The last scenario is the worst case where a leader leaves the group and there is no backup available. The base station needs to select a user from the affected member cluster and assign her/him as the new leader. Instead of rekeying the affected member cluster immediately, the KEK update is delayed until join, leave, or eventual departure of the chosen user takes place. The base station invokes the rekeying only in the leader cluster in order to update the group key. For instance, as shown in the Figure 2, when u1, u2, and

u3 leave the group, u5 is selected and assigned as the new leader of member cluster 1.

bs → u5: {new k_1, k_2, cluster-1 key, group key}k_{11}

bs → u4: {new k_1, k_3, group key}k_7

bs ⇒ cluster 1: {new cluster-1 key, group key}k_8

u4 ⇒ cluster 2: {new group key}$k_{(\text{cluster-2})}$

Handoff

Handoff is a unique operation in the wireless group communications. Several approaches (DeCleene et al., 2001; Sun et al., 2002) have been proposed in literature to address the group key management during the handoff. DeCleene et al. (2001) proposed a delayed rekeying scheme to postpone local rekeying until a particular criterion is satisfied, such as join, leave, or eventual departure of handoff users. We enhance this approach with the handoff authentication to further reduce the rekeying cost.

We assume that the mobile devices can detect the signals from two base stations on the edge of wireless cells. When a user moves from one cell to another at the edge of the cell, s/he switches to the new base station, sends a handoff join message, and then switches back to the old base station. The user authentication is performed by the two base stations (Wang & Le, 2005). The new base station sends a user authentication request to the old one:

$\text{bs}_{\text{new}} \rightarrow \text{bs}_{\text{old}}$: {AUTHENTICATION_REQUEST}

The old base station replies to the new base station with the authentication information of the moving user:

$\text{bs}_{\text{old}} \rightarrow \text{bs}_{\text{new}}$: {AUTHENTICATION_REPLY}

After a predefined time, the user switches to the new base station. When the user moves into the new cell, the new base station assigns the handoff user into a suitable cluster according to the join procedure, and sends a set of new keys to the user.

Instead of immediately rekeying, the old base station records the handoff user into a *Handoff User List* and postpones the rekeying until join or leaving happens in the cell or eventual departure of the moving user. When the local rekeying affects the handoff user, the base station deletes the user from the *Handoff User List.*

Performance Discussion

To measure the performance of a system, many parameters can be taken into consideration. However, we believe that for the group key management system, efficiency is one of the most significant parameters. In this section, we describe the efficiency of our proposed work in comparison with LKH. We explain the efficiency from three perspectives: communication cost, computation cost, and storage cost.

Communication Cost

Communication overhead is recorded as a measure of the number of rekeying messages transmitted per operation by the key server. We evaluate the communication cost for the join and leave operations. Without any loss of generality, we employ a binary tree to build the key tree.

Communication Cost During Join

In our structure, when a user joins the group, the new member can be either assigned into a leader cluster or member cluster. For the leader cluster join, the key update is restricted only to the leader cluster. Therefore, the communication cost is $\log_2 n_l$; n_l is the number of users in the leader cluster. As for the member cluster join, the communication cost is $\log_2 n_m + 2$; n_m is the number of users in the member cluster. According to LKH, the communication cost of join is $\log_2 n$, where n is the number of group users in the cell. In order to better explain the efficiency of our work, we calculate the join cost in one cell comprising 1,023 members. According to the cluster policy, the ratio of leaders and backups is 1:1, the members are divided into 17 clusters, 15 clusters having 64 members each and 2 clusters with 32 users. One 32-user cluster is assigned as the leader cluster. Now, we consider the join cost of a new member, namely 1024th user. In our approach, when the user joins the leader cluster, the cost is $\log_2 32 = 5$. If the user joins the member cluster, the overhead is $\log_2 64 + 2 = 8$. As for the LKH, the cost is $\log_2 1024 = 10$. So by comparison, we can see that our proposal achieves the significant improvement in the communication cost by 20% for member cluster join, and even better by 50% for leader cluster join.

Communication Cost During Leave

There are two types of leaving cost in our proposal: (1) user departure from member cluster, and (2) user departure from leader cluster. As for the user leaving from the member cluster, the communication cost is $\log_2 n_m + 2$, the summation of the rekeying cost of the member cluster key tree, and two extra messages sent to the leader cluster.

As for the user departing from the leader cluster, there are three situations: (1) backup user leaving, (2) cluster leader leaving with backup available, and (3) cluster leader leaving with no backup available. We calculate the communication cost of these three scenarios as follows:

Table 1. Comparison of communication cost

	Member Join	Member Leaving
Our Proposal	$\log_2 n_l$ / $\log_2 n_m + 2$	$\log_2 n_m + 2$ / $\log_2 n_l$ / $\log_2 n_l + 2$ / $\log_2 n_l + 2$
LKH	$\log_2 n$	$\log_2 n$

n: the number of group members in the cell
n_m: the number of users in the member cluster
n_l: the number of users in the leader cluster

When backup user leaves the group, the rekeying procedure is confined in the leader cluster. The communication cost is $\log_2 n_l$, where n_l is the number of users in the leader cluster. For the second situation, when the current cluster leader leaves the group, the backup can take over the leadership, so the communication cost is $\log_2 n_l + 2$. As for the third scenario where the leader and the backups move from the group, due to the delayed rekeying in the affected member cluster, the rekeying procedure is invoked only in the leader cluster. Hence, the communication cost is $\log_2 n_l + 2$.

In LKH, the leaving communication cost is $\log_2 n$, where n is the number of group members in the cell.

We consider the same example as described above to explain the situation further. In our proposal, when the user of the member cluster leaves, the communication cost is $\log_2 64 + 2 = 8$. When the user in the leader cluster leaves the group, the communication cost of three scenarios as described is $\log_2 32 = 5$, $\log_2 32 + 2 = 7$ and $\log_2 32 + 2 = 7$ respectively, whereas the rekeying cost for LKH is $\log_2 1024 = 10$. From this example, it can be observed that our proposal improves the communication efficiency by 30% and 50%.

We tabulate the comparison of our approach and LKH in Table 1.

Computation Cost

Computation cost is to measure the overhead of encryption and decryption during the rekeying procedure. This cost is firmly associated with the two factors: (1) the communication cost, and (2) the length of the message. In our approach, we group users into clusters and for each cluster a smaller key tree is built. Hence, during the rekeying procedure, the number of keys that needs to be updated is much less when compared to that of LKH, which means that the length of our rekeying message is much shorter than the message used in LKH. Additionally, from the above analysis, we can see that our proposal has a greater advantage over LKH with respect to the communication cost. By combining these two factors, it can be inferred that the proposed approach is much more computationally efficient than LKH.

Storage Cost

The storage efficiency is to measure the number of keys stored in the key server and at the user side. In our proposal, users are associated with the cluster key tree whose size is much smaller than the LKH key tree. Therefore our approach has less number of keys stored at the user side than that in LKH. For the users in the member cluster, the number of stored keys is $\log_2 n_m + 3$, and for the users in the leader cluster, the number of stored keys is $\log_2 n_l + 3$, whereas in LKH each user needs to store $\log_2 n + 1$ keys. Considering the same example, the user in the leader cluster needs to store 8 keys, and the number of keys stored by the user in the member cluster is 9. When compared to the 11 keys stored by the user in LKH, our approach proves to offer better efficiency. On the key server side, the number of keys stored in our proposal and LKH is $n_c \times 2n_m + 2n_l \approx 2n$ and $2n - 1$ respectively. We tabulate the comparison in Table 2.

FUTURE TRENDS

Along with the fast development in wireless technology and mobile devices, more and more multicast group applications and services will be emerging on wireless networks. The future research will focus on combining the proposed

Table 2. Comparison of key storage cost

	User	Key Server
Our Proposal	$\log_2 n_m + 3$ / $\log_2 n_l + 3$	$n_c \times 2n_m + 2n_l$
LKH	$\log_2 n + 1$	$2n - 1$

n_m: the number of users in the member cluster
n_l: the number of users in the leader cluster
n_c: the number of member clusters
n: the number of group members in the cell

work with other key management structures to provide a solution to secure and efficient group key management in wireless networks.

CONCLUSION

Here, we proposed a new, efficient group key management approach for wireless networks. The proposed scheme has a two-tier logical key structure where the users are grouped into clusters, which help in significantly improving the communication, computation, and storage efficiency on the client side.

REFERENCES

Chen, J., & Chao, T. (2004). *IP-based next-generation wireless networks.* New York: John Wiley & Sons.

Challal, Y., & Seba, H. (2005). Group key management protocols: A novel taxonomy. *International Journal of Information Technology, 2*(1), 105-118.

DeCleene, B., Dondeti, L. R., Griffin, S., Hardjono, T., Kiwior, D., Kurose, J., Towsley, D., Vasudevan, S., & Zhang, C. (2001, June). Secure group communications for wireless networks. *Proceedings of MILCOM.*

Hardjono, T., Cain, B., & Monga, I. (2000). *Intra-domain group key management protocol.* Retrieved from draft-ietf-ipsec-intragkm-00.txt

Harney, H., & Muckenhirn, C. (1997). Group key management protocol (GKMP) architecture. *RFC, 2094.*

Kim, Y., Perrig, A., & Tsudik, G. (2004). Tree-based group key agreement. *ACM Transactions on Information and System Security, 7,* 60-96.

Kostas, T., Kiwior, D., Rajappan, G., & Dalal, M. (2003). Key management for secure multicast group communication in mobile networks. *Proceedings of the DARPA Information Survivability Conference and Exposition* (Vol. 2, pp. 41-43).

Mittra, S. (1997). Iolus: A framework for scalable secure multicasting. *ACM SIGCOMM, 27*(4).

Perrig, A., Song, D., & Tygar, D. (2001). ELK, a new protocol for efficient large-group key distribution. *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 247-262).

Sherman, A. T., & McGrew, D. A. (2003). Key establishment in large dynamic groups using one-way function trees. *IEEE Transactions on Software Engineering, 29,* 444-458.

Steiner, M., Tsudik, G., & Waidner, M. (1996). Diffie-Hellman key distribution extended to group communication. *Proceedings of the 3rd ACM Conference on Computer and Communications Security (CCS '96)* (pp. 31-37).

Sun, Y., Trappe, W., & Liu, K. J. R. (2002). An efficient key management scheme for secure wireless multicast. *Proceedings of the IEEE International Conference on Communications* (Vol. 2, pp. 1236-1240).

Wallner, D., Harder, E., & Agee, R. (1999). Key management for multicast: Issues and Architectures. *RFC, 2627.*

Wang, YL., & Le, P.D. (2005, September). Secure group communications in wireless networks. *Proceedings of the 3rd International Conference on Advances in Mobile Multimedia,* Malaysia.

Wong, C. K., Gouda, M., & Lam, S. S. (2000). Secure group communications using key graphs. *IEEE/ACM Transactions on Networking, 8,* 16-30.

KEY TERMS

Backward Secrecy: To prevent new group members from accessing previous group communications, which they may have recorded.

Forward Secrecy: To prevent departing members from decoding future group data traffic.

Handoff: In a cellular wireless network, the transition of signal for any given user from one base station to a geographically adjacent base station as the user moves around.

Key Encryption Key (KEK): A key used to encrypt the other keys for distribution in the multicast group.

Key Management Algorithm: In the group key management system, an algorithm is applied to maintain the logical key structure held by the group members and other entities.

Logical Key Hierarchy (LKH): This type of algorithm is a tree structure for efficient group rekeying. Each node of the tree represents a key, with the root node representing the group key. Each leaf node represents a group member, and each member knows all the keys in its path to the root.

Multicast: A communication mechanism to delivery a single message to multiple receivers on a network. The message will be duplicated automatically by routers when multiple copies are needed.

1-Affects-n: When one group membership changes, the rekeying procedure will affect all the remaining members.

Efficient Replication Management Techniques for Mobile Databases

Ziyad Tariq Abdul-Mehdi
Multimedia University, Malaysia

Ali Bin Mamat
Universiti Putra Malaysia, Malaysia

Hamidah Ibrahim
Universiti Putra Malaysia, Malaysia

Mustafa M. Dirs
College University Technology Tun Hussein Onn, Malaysia

INTRODUCTION

Mobile databases permeate everywhere into today's computing and communication environment. One envisions application infrastructures that will increasingly rely on mobile technology. Current mobility applications tend to have a large central server and use mobile platforms only as caching devices. We want to elevate the role of mobile computers to first class entities in the sense that they will allow the mobile user to work/update capabilities independent of a central server. In such an environment, several mobile computers may collectively form the entire distributed system of interest. These mobile computers may communicate to each other in an ad hoc manner by communicating through networks that are formed on demand. Such communication may occur through wired (fixed) or wireless (ad hoc) networks. At any given time, a subset of the computer collection may connect and would require reliable and dependable access to relevant data of interest. Peer-to-peer (P2P) computing, basically, is an ad hoc network and it can be built on the fixed or along a wireless network. With P2P, computers can communicate directly and share both data and resources. So far, many applications such as ICQ (where users exchange personal messages), similar to Napster and Freenet (where users exchange music files), have taken the advantage of P2P technology. However, data management is an outstanding issue and leads directly to the problem of low data availability. Data availability is the central issue in P2P data management. The most important characteristic that affects data availability in P2P environment is the nature of the network. In the case of an ad hoc network, hosts are connected to the network only temporarily. Furthermore, hosts play the role of router, and they communicate with each other directly without any dedicated hosts. If there are no dedicated hosts that act as a router, obviously the network connections are prone to get disconnected and/or become unreliable. Thus, it is difficult to guarantee one-copy "serializability," since one relies on the mobile hosts, not the fixed hosts, in order to communicate with other hosts not reachable directly (Faiz & Zaslavsky, 1995). When hosts disconnect more often, due to the applications that have high transaction rates, the deadlock and reconciliation rate will experience a cubic growth (Faiz & Zaslavsky, 1995) and, the database is in an inconsistent state and there is no obvious way to repair this problem or allow for this eventuality. In the case of fixed network, the network connection is relatively stable, but the availability of sufficient computing resource depends on the strategies of replication.

Walborn and Chrysanthis (1997) describe the use of mobile computers in the trucking industry. Each truck has a computer with a satellite/ radio link—this is able to interact with the corporate database. Other applications include involving avoiding remote or disaster areas and for military applications with mobile computers forming ad hoc networks without communications and/or with stationary computers. Faiz and Zaslavsky (1995) discuss the impact of wireless technologies and mobile hosts on a variety of replication strategies. Distributed replicated file systems such as Ficus and Coda (Reiher, Heidemann, Ratner, Skinner, & Popek, 1994) have extensive experience with disconnected operations.

In this article, we consider the distributed database that can make up mobile nodes as well as peer-to-peer concepts. These nodes and peers may be replicated both for fault tolerance (dependability), and to compensate for nodes that are currently disconnected. Thus we have a distributed replicated database, where several sites must participate in the synchronization of a transaction. The capabilities of the distributed replicated database are extended to allow mobile nodes to plan for a disconnection, with the capability of update, and for the database—on behalf of mobile node by using fixed proxy server—to make these updates during the mobile

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

disconnection. Once a mobile reconnects, it automatically synchronizes and integrates into database.

By using the notion of planned disconnection (*sign-off and check-out modes*), we present a framework, which allows the replicated data of mobile nodes to be available to access and update for low costs in reading and writing.

This article is organized as follows; in the second section, we review the *read one write all* (ROWA) technique; in the third section, the model of the *diagonal replication on grid* (DRG) technique is presented, and we also present an algorithm to allow disconnection nodes to update using a sign-off and check–out idea adapted in the system. In the fourth section, the correctness and the performance of the proposed technique is analyzed in terms of *communication cost and availability* comparing ROWA techniques and DRG techniques. In the final section, the conclusion is given.

VIEW OF ROWA STRUCTURE TECHNIQUE

The simplest technique to maintain replicated data is when a read-only operation is allowed to read any copy, and write operation is required to write all copies. This is called a read one write all (ROWA) protocol. This protocol only works correctly when a transaction process from one correct state to another correct state is carried out. The ROWA has the lowest read cost because only one replica is accessed by a read operation. The weakness of this method is the low write availability, because a write operation cannot be done in a failure of any replica.

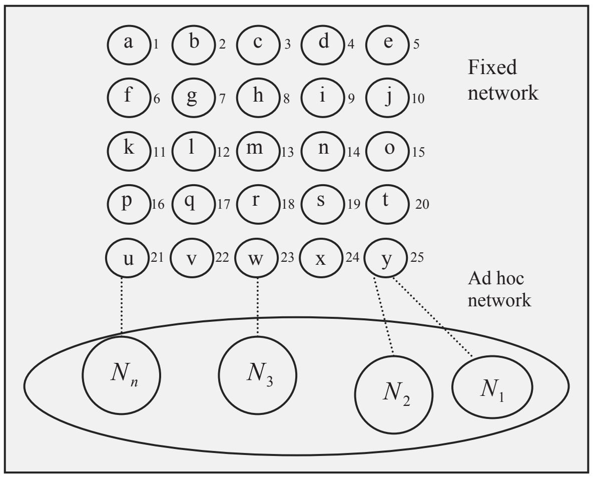
The available copies technique proposed by Bernstein, Hadzilacos, and Goodman (1987) is an enhanced version of the ROWA approach in terms of the availability of write operations. Every read is translated into a "read" of any replica of the data object and every "write" is translated into write of all available copies of that data object. This technique can handle each site either when it is operational or down—and that all operational sites can communicate with them. If a site does not respond to a message within the timeout period, then it is assumed to be down. However, writing is very expensive when all copies are available: forcing read-write transactions to write all replicas.

Lazy replication protocol does not attempt to perform the write operation on all copies of the data object within the context of the transaction that updates that data object. Instead, it performs the update on one or more copies of the data object and later propagates the changes to all the other copies in all other sites. A lazy replication scheme can be characterized using four basic parameters(Bernstein, Hadzilacos, & Goodman, 1987; Borowski, 1996; Goldring, 1995; Ozsu & Valduriez, 1999). The ownership parameter defines the permissions for updating copies. If a copy is updateable it is called primary copy, otherwise it is called a secondary copy. The site that stores the primary copy of a data object is called a master for this data object, while the sites that store its secondary copies are called slaves. The propagation parameter defines when the updates copy must be propagated towards the sites storing the other copies of the same data object. Generally, lazy replication protocols can be classified into two groups. (1) The first group consists of lazy replication methods where all copies are updateable. In this case, there is group ownership on the copies. A conflict happens if two or more sites update the same replica. There are several policies for conflict detection and resolution that can be based on timestamp ordering, node priority, and others (Buretta, 1997; Helal, Heddaya, & Bhargava, 1996). The problem with conflict resolution is that during a certain period of time the database may be in an inconsistent state. (2) The second group consists of protocols where there is a single master that is updated, and each time a query is submitted for execution, the secondary copies that are read by the query are refreshed by executing all received refresh transaction. Therefore, a delay may be caused by that query.

DATABASE MODEL

In considering this environment, it has two types of networks; that is, a fixed network or an ad hoc network. For the fixed network, all sites are logically organized in the form of two-dimensional grid structure (i.e., 5*5) and in a nodes (ad hoc network), this consists of N nodes labeled $N_1, N_2, \ldots, N_n$ (as shown in Figure 1). The data will only be partially or fully replicated in that data items are stored redundantly at multiple sites. Information about the location of all copies of a data item may be stored at each site or kept in directories at several of the sites. Users interact with the database by invoking transactions at any one of the database sites. A

Figure 1. Database model



transaction is a sequence of read and writes operations on the data items that are executed automatically. The criterion for correctness in databases is the *serializable* execution of transactions (Wolfson, Jojodia, & Huang, 1997). Serializable executions are guaranteed by using a *concurrency control* mechanism, such as two-phase locking, timestamp ordering, or optimistic concurrency control (Agrawal & Abbadi, 1996). Since two-phase locking is widely used, we assume that each site in the distributed system enforces two-phase locking locally. Multiple copies of a data object must appear as a single logical data object to the transactions. This is termed as "one-copy equivalence" and is enforced by the replica control technique; as used in this study of diagonal replication grid (DRG) technique. The correctness criterion for replicated database is one-copy serializability (Berstein & Goodman, 1994), which ensures both one-copy equivalence and the serializable execution of transactions. In order to ensure one-copy serializability, a replicated data object may be read by reading a quorum of copies, and it may be written by writing a quorum of copies, The selection of a quorum is restricted by the quorum intersection property to ensure one-copy equivalence: for any two operations o [x] and o' [x] on a data object x, where at least one of them is a write, the quorum must have a non-empty intersection. The quorum for an operation is defined as a set of copies whose number is sufficient to execute that operation. We assume that a mechanism to enforce one-copy serializability is used. This could be a "synchronous control protocol" using diagonal replication on grid (DRG) technique, that is, a site *initiates* a DRG transaction to update its data object. For all accessible data objects, a DRG transaction attempts to access a DRG quorum. If a DRG transaction gets a DRG write quorum without non-empty intersection, it is accepted for execution and completion, otherwise it is rejected. We assume that for a read quorum, that if two transactions attempt to read common data objects, read operations do not change the values of the data object. Since read and write quorums must intersect and any two DRG quorums must also intersect, then all transaction executions are one-copy serializable.

In this article, we consider two models—*sign-off model and check- out model*. They work under these conditions:

1. The distributed system must be able to process database updates even though some of the nodes are not available. It is to designate one of the members as the fixed proxy site at fixed network, which can present mobile node when mobile disconnection.
2. To optimize the communication costs and the availability of the system of replicated data in the distributed systems under the fixed and ad-hoc network in P2P environment.
3. The fixed network, we are describing in the diagonal replication on grid (DRG) technique, considers only that the diagonal sides will be replicated—that of the data at a fixed network based on quorum intersection property.
4. The members of mobile nodes are fixed and databases at ad-hoc network (mobile nodes) will be replicated at the node that is the most commonly visited node.
5. The fixed proxy should be selected from the nearest one to the mobile database.
6. Transfer any data between two mobiles nodes should be through fixed proxy.

Basic Sign-off

In a simple planned disconnection or *basic sign-off* (Holliday, Agrawa, & Abbadi, 2000), the database of the disconnected node becomes read-only. The connected sites continue to read and update the data item or objects. DRG technique adjusts the replication at diagonal sites at fixed network so that the transaction can complete in spite of mobile node disconnects. The planned disconnection will be accomplished with the help of a fixed *prox*y. When a node disconnects, it appoints another site as a fixed network to vote on its behalf to ensure that replicas can be updated, and in any other actions of the distributed system that require consensus. The power to vote on behalf of a mobile node at fixed network is called a fixed proxy and the site with that power is also referred to as a fixed proxy.

If node N_1 wants to disconnect from a distributed database, it carries out a disconnect dialog so that the system is aware that it has not failed, but will merely disconnect for a period of time. The node N_1 contacts the nearest fixed network Y to be N_1's fixed proxy. During the disconnection, N_1 can only read its local copy of the data. When the fixed proxy Y sees a message for N_1, it answers on behalf of N_1 while N_1 is disconnected. The fixed proxy Y also keeps track of the updates to the database that N_1 has missed because of the disconnection. (We assume that all updates are sent to all copies in a diagonal set, and so, fixed proxy Y can respond on behalf on N_1).

Assuming that N_1 is a mobile computer and wishes to disconnect using sign-off procedure:

1. Node N_1 selects the nearest fixed proxy Y from fixed network, the fixed proxy Y should be peer to N_1. The fixed proxy Y should inform N_1 which will be the fixed proxy to N_1 in case of fixed proxy Y does not work (failure).
2. Node N_1 informs fixed proxy Y the new data that is not replicated at fixed proxy Y, N_1 transfer a new data.
3. The fixed proxy Y will replicate the N_1 data to all diagonal set on the request to it.
4. The data items at N_1 that are replicated at fixed proxy Y, the fixed proxy Y has the right to vote for N_1 in matters concerning writes to the data items.

When node N_1 wants to reconnect, it should also go through sign-on:

1. Node N_1 reconnects and contacts the fixed proxy Y.
2. Fixed proxy Y transfers all new data that has been missed during disconnection to node N_1.
3. If the fixed proxy Y itself disconnects during system failure, the other site that is appointed from fixed proxy Y can work as fixed proxy to node N_1.
4. During connection of node N_1, the fixed proxy Y (or the one on behalf of the fixed proxy Y) allows services to transfer data anew to node N_1 —all which has been missed during the disconnection of node N_1.

In case of node N_1 being disconnected forever, the node N_1 will inform fixed proxy Y whether to delete all N_1's data or keep it with fixed proxy Y forever, or node N_1 will leave their data at fixed proxy Y when planned disconnect and forget it, this will depend on node N_1.

Check-Out Model

If the node N_1 wants to disconnect and still be able to update a particular data object, it declares its intention to do so, but before disconnection—and "check-out" or "takes" the object for writing. This is accomplished by obtaining a lock on the item before disconnection. In order to maintain serializability in *check-out* mode, the fixed proxy Y and its diagonal set including the mobile nodes are prevented from accessing the object which N_1 has checked-out (as if N_1 had a write-lock). An object can only be checked-out to one mobile node at a time. Since many database systems use two-phase locking, it makes sense to implement *check-out* mode using the existing locking mechanisms. The mobile node that wishes to disconnect, for example, N_1, acquires a write-lock on the item or object it wants to update while disconnected. This write-lock is like an ordinary write-lock except that the "transaction" that holds it should not be aborted due to a deadlock with ordinary transactions. The mechanism for obtaining the lock might be via a transaction or through some other means. In order to distinguish these "transactions" from ordinary user transactions, we will call them *pseudo-transactions*.

To preserve correctness, it must be possible to serialize all of the transactions executed by node N_1 during disconnection and at the point in time of disconnection. This can be done if:

1. Only those items write-locked by pseudo-transactions at disconnect time can be modified by node N_1 during disconnect.
2. Items write-locked by pseudo-transactions at disconnect time can neither be read or written by other sites (a consequence of maintaining the write-lock) and the pseudo-transaction cannot abort in order to release the lock.
3. Items not write-locked by pseudo-transactions at disconnect time are treated as read-only by node N_1 during disconnect (unless they were currently locked by other transactions at the time of disconnection).

Assuming that node N_1 wishes to disconnect and "check-out" a set of items—X. The disconnect procedure is as follows:

1. Node N_1 selects a fixed proxy—Y, as in basic sign-off mode—and follows all the same steps to handle voting rights for replicated and non-replicated items.
2. At the same time, node N_1 initiates a pseudo-transaction to obtain write-lock on the items in X.
3. If the pseudo-transaction is successful, N_1 disconnects with update privileges on all the items in X. If the pseudo-transaction is not successful, N_1 will try again or disconnect without update rights to X.

When node N_1 wants to reconnect, it should also go through check-out:

1. Node N_1 reconnects and contacts the fixed proxy—Y.
2. Node N_1 will transfer any new data from X to fixed proxy Y.
3. Fixed proxy Y will commit the value of item X and release the lock from item X.
4. Fixed proxy Y transfers any new data that has been missed during disconnection to node N_1.
5. If the fixed proxy Y itself disconnects (e.g., due system failure), the other site that has been appointed from fixed proxy Y can work as fixed proxy to node N_1.

DRG Technique

This environment has two types of networks, (1) the fixed network and (2) the ad-hoc network. For the fixed network, all sites are logically organized in the form of a two-dimensional grid structure. For example, if a DRG consists of twenty-five sites, it will logically organized in the form of 5 x 5 grid (as shown in Figure 1), each site having a master data file. In the remainder of this study, we are assuming that all replica copies are data files. A site is either operational or failed, and the state (operational or failed) of each site is statistically independent to the others. When a site is operational, the copy at the site is available; otherwise it is unavailable. In the fixed network, the data file will replicate to *diagonal sites*. While in the ad-hoc network, the data file will replicate asynchronously at only one node based on the most frequently visited site (when the node reconnects the fixed proxy will transfer to it any new and recent updated data). The logical structure for fixed and ad hoc network is shown as in Figure 1. The circles in the grid represent the sites under the fixed network environment and *a,b,…,y* represent the master data files located at site 1,2,…,25 respectively. The circles S1,S2,…., Sn represent the master file at mobile node located at node 26,27,28,…,n; as shown in the oblong shape are nodes under the ad-hoc network.

The commonly visited site is defined as the most frequented node request for the same data at a fixed network (the commonly visited sites can be given either by a user or selected automatically from a log file/database at each center). This site will replicate the data asynchronously (by/once mobile node reconnects to a fixed proxy), until then it will not be considered for read and write quorums on fixed network, but mobile nodes can read their own data during disconnection without any update of the data. Since the data file is replicated to only the diagonal sites at the fixed network, therefore it minimizes the number of database update operations, misrouted (and dropped out calls). Also, sites are autonomous in processing different queries or update operations; this consequently reduces the query response time. The number of data replication, d, can be calculated using Property 1, described as follows:

Property 1. One assumes the number of data replication from each site, $d = n$

Proof: Let N = n x n be a set of all sites that are logically organized in a two-dimensional grid structure form as shown in Figure 1. Based on definition 1:

The number of diagonal sites = number of sites in a diagonal set

$= | D(s) | \; = n.$
$\because\ n=5$
$\because\ n=|D(s)|$
$\therefore |D(s)|=5$

Definition 1. Assuming that the fixed network environment consists of n x n sites that they are logically organized in the form of a two-dimensional grid structure. These sites are labeled $s(i,j)$, $1\le i\le n$,$1\le j \le n$. The *diagonal site* to $s(i,j)$ is $\{s(k,l)|\ k=i+1,\ l = j+1$; and $\{k,l \le n$,if i=n, initialized i=0,if j=n, initialized j=0$\}$. A diagonal set, D(s), is a set of diagonal sites. As an example, Assume that n=5, then the diagonal site to s(1,1) is s(2,2), the diagonal site to s(2,2) is s(3,3), the diagonal site to s(2,1) is (3,2), and so forth.

Thus, based on this technique, sites in the diagonal set will have the replica copies in common. From Figure 1, one of the diagonal sets is {s(1,1), s(2,2), s(3,3), s(4,4),s(5,5)}, and each site will have the same replica copies, that is, {a,g,m,s,y}.

The number of diagonal set equals to n, and the m^th^ diagonal set is noted as $D^m(s)$, for m=1,2,…n.

For example, from *Figure 1*, if n=5, then the diagonal sets are:

$D^1(s)$ = {s(1,1), s(2,2), s(3,3), s(4,4),s(5,5)},
$D^2(s)$ = {s(2,1), s(3,2), s(4,3), s(5,4),s(1,5)},
$D^3(s)$ = {s(3,1), s(4,2), s(5,3), s(1,4),s(2,5)},
$D^4(s)$ = {s(4,1), s(5,2), s(1,3), s(2,4),s(3,5)}, and
$D^5(s)$ = {s(5,1), s(1,2), s(2,3), s(3,4),s(4,5)},

The primary site of any data file and, for simplicity, its diagonal sites, are assigned with vote one and vote zero, which is analogous to binary vote assignment proposed in (Mat Deris, Evans, Saman, & Noraziah, 2000). A vote assignment on grid, B, is a function such that

$$B(s(i,j)) \in \{0,1\},\ 1\le i \le n, 1\le j \le n$$

where $B(s(i,j))$ is the vote assigned to site $s(i,j)$. This assignment is treated as an allocation of replicated copies and a vote assigned to the site results in a copy allocated at the diagonal site. That is, 1 vote ≡ 1 copy.

Let

$$L_B = \sum_{s(i,j)\in D(s)} B(s(i,j))$$

where, L_B is the total number of votes assigned to the primary site and its diagonal sites. Thus, $L_B = d$.

Let r and w denote the read quorum and write quorum, respectively. To ensure that the read operation always gets up-to-date values, r + w must be greater than the total number of copies (votes) assigned to all sites. The following conditions are used to ensure consistency:

$$1\le r\le L_B,\ 1\le w\le L_B,$$

$$r + w = L_B + 1.$$

Conditions (1) and (2) ensure that there is a non-empty intersection of copies between every pair of read and write operations. Thus, the conditions ensure that a read operation can access the most recently updated copy of the replicated data. Timestamps can be used to determine which copies are most recently updated.

Let S(B) be the set of sites at which replicated copies are stored corresponding to the assignment B. Then

$$S(B) = \{s(i,j)|\ B(s(i,j)) = 1,\ 1\le i\le n,\ 1\le j\le n\}.$$

Definition 2. For a quorum q, a *quorum group* is any subset of S(B) whose size is greater than or equal to q. The collection of quorum group is defined as the *quorum set*.

Let Q(B,q) be the quorum set with respect to the assignment B and quorum q, then

$$Q(B,q) = \{G \mid G \subseteq S(B) \text{ and } |G| \geq q\}$$

For example, from *Figure 1*, let site s(1,1) be the primary site of the master data file *a*. Its diagonal sites are s(2,2),s(3,3)),s(4,4), and s(5,5). Consider an assignment B for the data file *a*, such that

$$B_a(s(1,1))=B_a(s(2,2))=B_a(s(3,3))=B_a(s(4,4))=B_a(s(5,5)) = 1$$

and

$$L_{B_a} = B_a(s(1,1))+B_a(s(2,2))+B_a(s(3,3))+ B_a(s(4,4)) + B_a(s(5,5)) = 5.$$

Therefore, $S(B_a)$ = {s(1,1),s(2,2),s(3,3), s(4,4),s(5,5)}.

If a read quorum for data file *a*, r =2 and a write quorum w = L_{Ba}-r+1 = 4, then the quorum sets for read and write operations are Q(B_a,2) and Q(B_a,4), respectively, where

Q(B_a,2)={s(1,1),s(2,2)},{s(1,1),s(3,3)},{s(1,1),s(4,4)},{s(1,1),s(5,5)},{s(2,2),s(3,3)},{s(2,2),s(4,4)},{s(2,2),s(5,5)},{s(3,3),s(4,4)},{s(4,4),s(5,5)},{s(1,1),s(2,2),s(3,3)},

{s(1,1),s(2,2),s(4,4)},{s(1,1),s(2,2),s(5,5)},{s(1,1),s(3,3),s(4,4)},{s(1,1),s(3,3),s(5,5)},

[s(1,1),s(4,4),s(5,5)},{s(2,2),s(3,3),s(4,4)},{s(2,2),s(3,3),s(5,5)},{s(2,2),s(4,4),s(5,5)},

{s(3,3),s(4,4),s(5,5)},{s(1,1),s(2,2),s(3,3),s(4,4)},{s(1,1),s(2,2),s(3,3),s(5,5)}, {s(1,1),s(2,2),s(4,4),s(5,5)},{s(1,1),s(3,3),s(4,4),s(5,5)},{s(2,2),s(3,3),s(4,4),

s(5,5)},{s(1,1),s(2,2),s(3,3),s(4,4),s(5,5)}And

Q(B_a,4)={s(1,1),s(2,2),s(3,3),s(4,4)},{s(1,1),s(2,2),s(3,3),s(5,5)},{s(1,1),s(2,2),s(4,4),s(5,5)},

{s(1,1),s(3,3),s(4,4),s(5,5)},{s(2,2),s(3,3),s(4,4),s(5,5)},{s(1,1),s(2,2),s(3,3),s(4,4),s(5,5)}

The Correctness of DRG

In this section, the study will show that the DRG protocol is one-copy serializable. The sets of groups (coterie) (Maekawa, 1985) will be defined, and to avoid confusion we refer to sets of copies as groups. Thus, a set of groups are/is a set of sets of copies.

Definition 3. Coterie. Let U be a set of groups that compose the system. A set of groups T is a coterie under U if and only if:

1. *G Î T implies that G ¹ Æ and G Í U*
2. *G, H Î T then G Ç H ¹ Æ* (int*ersec*tion *property*).
3. *There are no G, H Î T such that G Ì H* (min*imality*)

By the definition of coterie and definition from 3.3.2, then Q (B, w) is a coterie, because it satisfies all coteries' properties. The correct criterion for replicated database is one-copy serializable. The next theorem provides us with a mechanism to check whether DRG is correct.

Assertion 3. The history H is one-copy serializable if Ti Î H, i=1,2,…n satisfy quorum intersection properties.

Proof: Suppose history H satisfies quorum intersection properties. Assume that Ti, Tj Î H, then at least one of Ti's operations precedes and conflicts one of Tj's operations. Then Ti→Tj. Thus H is an cyclic RDSG. By theorem (4) H is one- copy serializable.

Theorem 3. The DRG protocol is one-copy serializable.

Proof: as in Assertion 1. The theorem holds on condition that the DRG protocol satisfies the quorum intersection properties. Since read operations do not change the value of the accessed data object, a read quorum does not need to satisfy the intersection property. To ensure that a read operation can access the most recently updated copy of the replicated data, that means the two conditions as follow must be conformed.

1. $1 \leq r \leq L_B$, $1 \leq w \leq L_B$
2. $r + w = L_B + 1$

While a write quorum needs to satisfy read-write and write-write intersection properties. For case of write-write intersection, since W is coterie then it satisfies write-write intersection. However, for the case of read-write intersection, it can be easily shown that ∀ *G Î R and* ∀ *H Î W, then G Ç H ¹ Æ.*

PERFORMANCE ANALYSIS AND COMPARISON

In this section, we analyze and compare the performance of the DRG technique with the ROWA technique on the communication cost and the data availability.

Communication Costs and Availability Analysis

The communication cost of an operation is directly proportional to the size of the quorum required to execute the operation. Therefore, one represents the communication cost in terms of the quorum size. In estimating the availability of this, all copies are assumed to have the same availability *p*.

$C_{X,Y}$ denotes the communication cost with X technique for Y operation, which is R(read) or W(write).

The ROWA Technique

Let N be the number of copies which are organized as a dimension n x n. Read operation needs only one copy, while a write operation needs to access n copies (a copy in each replica) in the system. Thus, the communication cost of a read operation is:

$$C_{ROWA,R} = 1$$

and the communication cost of write operation is:

$$C_{ROWA,W} = n$$

In the case of quorum, ROWA requires a read on any one of the copies, therefore the availability for read operation in the ROWA technique is

$$\begin{aligned} A_{ROWA,R} &= \sum_{i=1}^{n} \binom{n}{i} P^{i} \left(1-P\right)^{n-i} \\ &= 1-\left(1-P\right)^{n}. \end{aligned} \qquad (1)$$

While in write operation it needs to writes in all copies. Let $A_{X,Y}$ be the availability with X technique for Y operation, then the write availability in the ROWA technique is:

$$\begin{aligned} A_{ROWA,W} &= \sum_{i=1}^{n} \binom{n}{i} P^{i} \left(1-P\right)^{n-i} \\ &= P^{n}. \end{aligned} \qquad \text{.....(2)}$$

The DRG Technique

Let p_i denote the availability of site i. Read operations on the replicated data are executed by acquiring a read quorum and write operations are executed by acquiring a write quorum. For simplicity, one chooses the read quorum equal to the write quorum. Thus, the communication cost for read and write operations equals to $\lfloor L_{Ba}/2 \rfloor$, that is,

$$C_{DRG,R} = C_{DRG,W} = \lfloor L_{Ba}/2 \rfloor.$$

For example, if the primary site has four neighbors, each of which has vote one, then

$$C_{DRG,R} = C_{DRG,W} = \lfloor 5/2 \rfloor = 3.$$

For any assignment B and quorum q for the data file x, define $\varphi(B_a,q)$ to be the probability that at least q sites in $S(B_a)$ are available, then:

$$\varphi(B_a,q) = \Pr\{\text{at least q sites in } S(B_a) \text{ are available}\}$$

$$= \sum_{q}^{n} \binom{|S(B_x|}{q} p^{q} (1-p)^{|S(B_x|-q} \qquad \text{..... (3)}$$

Thus, the availability of read and write operations for the data file *a*, are $\varphi(B_a,r)$ and $\varphi(B_a,w)$, respectively. Let $Av(B_a,r,w)$ denote the read/write availability corresponding to the assignment B_a, read quorum r and write quorum w. If the probability that an arriving operation of read and write for data file *a* are *f* and (1-*f*), respectively, then:

$$Av(B_a,r,w) = f\,\varphi(B_a,r) + (1-f)\,\varphi(B_a,w). \qquad \text{...... (4)}$$

Definition 4. Let Av(x) be the availability function with respect to x. Av(x) is in the closed form if $0 \le x \le 1$ then $0 \le Av(x) \le 1$.

Theorem 4. The read/write availability under DRG technique are in the closed form.

Proof: For the case of read availability, from equation (3) and by definition 3.1.2, as $0 \le p_i \le 1$, $i=1,2,\ldots,L_{Ba}$ then $0 \le \varphi(B_a,r) \le 1$. Similarly, for the case of write availability where $0 \le \varphi(B_a,w) \le 1$ as $0 \le p_i \le 1$

Performance Comparison

Comparison of Costs

The communication cost of an operation is directly proportional to the size of the quorum required to execute the operation. Therefore, one represents the communication cost in terms of the quorum size. Table 1 shows the read and write costs of the two techniques between DRG and ROWA for different total number of copies, n = 16, 25, 36, 49, 64, and 81. The compared data of ROWA is derived from MUSTAFA. From Table 1, it is apparent that DRG has the lowest cost for write operations in spite of having a bigger

Table 1. Comparison of the read and write cost between GS and DRG under the different copies

	Number of copies in the system					
	16	25	36	49	64	81
ROWA (R)	1	1	1	1	1	1
ROWA (W)	16	25	36	49	64	81
DRG(R)	2	3	3	4	4	5
DRG (W)	3	3	4	4	5	5

Table 2. Comparison of the read availability between ROWA and DRG

Techniques	Read Availability								
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
ROWA, N=25	1	1	1	1	1	1	1	1	1
ROWA, N=36	1	1	1	1	1	1	1	1	1
ROWA, N=64	1	1	1	1	1	1	1	1	1
DRG, N=25	0.837	0.837	0.837	0.837	0.837	0.837	0.837	0.837	0.837
DRG, N=36	0.763	0.781	0.8	0.818	0.837	0.855	0.874	0.892	0.911
DRG, N=64	0.82	0.833	0.847	0.86	0.874	0.888	0.901	0.915	0.928

Table 3. Comparison of the write availability between ROWA and DRG

Techniques	Write Availability								
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
ROWA, N=25	0	0	0	0	0	0	0	0	0
ROWA, N=36	0	0	0	0	0	0	0	0	0
ROWA, N=64	0	0	0	0	0	0	0	0	0
DRG, N=25	0.837	0.837	0.837	0.837	0.837	0.837	0.837	0.837	0.837
DRG, N=36	0.763	0.781	0.8	0.818	0.837	0.855	0.874	0.892	0.911
DRG, N=64	0.82	0.833	0.847	0.86	0.874	0.888	0.901	0.915	0.928

number of copies when compared with ROWA quorums. It can be seen that DRG needs only 3 copies for the write quorums with 25 copies. On the opposite, the write cost is 25 for the ROWA with 25 copies. It increases more than DRG as a number of copies increases. For example, it increases to 81 for the write cost for ROWA with 81 copies, while for DRG; it increases to 5 for the write cost. While in the read operation it is apparent that ROWA has the lowest cost for read operations in spite of having a bigger number of copies when compared with DRG quorums because the data object will replicate at all the sites and (in this case) will increase the accessibility and availability of the data and make the cost of reading equal to 1 compared with a DRG cost between 2 and 5 for copies between 16 and 81. Looking at the lowest cost or reading in the ROWA it is still not efficient because the copy of the database is not replicated at all the sites causing a disadvantage in that this take large space for storage with an increase in the response time. So still the read at DRG technique is better than ROWA because the data item will replicated to diagonal sites only and this case will decrease the response time and updating.

Figure 2. Comparison of the read availability between ROWA and DRG

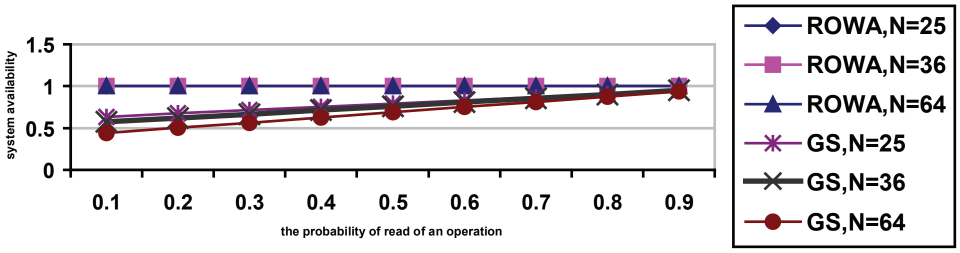
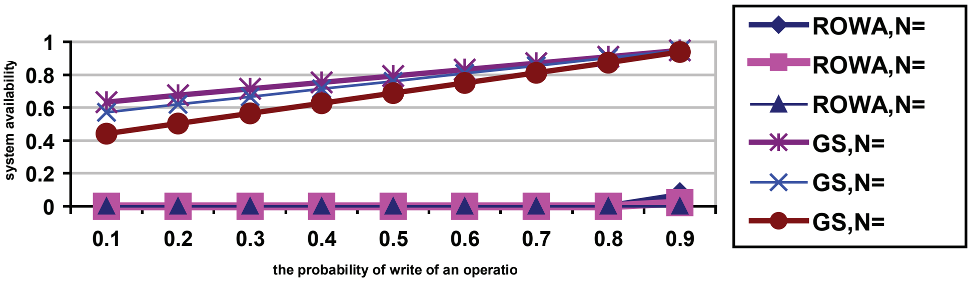


Figure 3. Comparison of the read availability between ROWA and DRG



Comparisons of Read/Write Availabilities

In this section, one will compare the performance on the read/write availability of the ROWA technique based on equations (1) and (2), and our DRG technique based on equations (3) and (4) for the case of n= 25, 36, and 64. In estimating the availability of operations, all copies are assumed to have the same availability.

Figures 2 and 3 and Tables 2 and 3 show the results obtained from the analysis for read and write availabilities between those two protocols when n=25,36 and 64.We assume that all data copies have the same availability, p, and varies from 0.1 to 0.9. From Tables 2 and 3, note that read availability for ROWA outperformed the DRG technique. This is due to the fact that ROWA needs only one copy for the communication cost. However those copies have the availability of more than 90% when individual copy has the availability of 70% and above. On opposite, the write availability for DRG outperformed from ROWA. This is due to fact that the number of copies needed to construct.

CONCLUSION

In this article we show the typical two-planned disconnection models to be suitable with P2P system, but we have presented a new technique, called the "Hybrid Replication Technique." This has been proposed to manage data replication in the fixed and ad hoc P2P network environment. The replication technique for the fixed network is proposed and based on diagonal replication technique (DRG), while the replication for the ad hoc network is done asynchronously only to the most frequently visited site by using one of the planned disconnection models. The analysis of the DRG has been presented in terms of read/write availability and communication costs. This has showed that, the DRG technique provides a convenient approach to high availability for update-frequent operations. This is due to the minimum number of quorum size required. In comparison to the ROWA, DRG requires significantly lower communication cost for an operation, while providing higher system availability, which is preferred for P2P environment.

REFERENCES

Agrawal, D., & El Abbadi, A. (1996). Using reconfiguration for efficient management of replicated data. *IEEE Transactions on Knowledge and Data Engineering*, *8*(5), 786-801.

Bernstein, P. A., Hadzilacos, V., & Goodman, N. (1987). Concurrency control and recovery in database systems. Reading, MA: Addison Wesley.

Bernstein, P. A., & Goodman, N. (1994). An algorithm for concurrency control and recovery in replicated distributed databases. *ACM Transactions on Database Systems*, *9*(4), 596-615.

Borowski, S. (1996). *Oracle 7 Concepts Release 7.3*. Redwood City, CA: Oracle Corp.

Buretta, M. (1997). Data replication: Tools and techniques for managing distributed information. New York: John Wiley.

Faiz, M., & Zaslavsky, A. (1995, March). Database replica management strategies in multidatabase systems with mobile hosts. In *Proceedings of the 6th International Hong Kong Computer Society Database Workshop*.

Goldring, R. (1995, May). Things very update replication customer should know. In *Proceedings of ACM SIGMOD International Conference On Management of Data* (pp. 439-440).

Helal, A. A., Heddaya, A. A., & Bhargava, B. B. (1996). *Replication techniques in distributed systems*. MA: Kluwer Academic Publishers.

Holliday, J., Agrawal, D., & Abbadi, A. E. (2000, February). Exploiting planned disconnections in mobile environments. In *Proceedings of the 10th IEEE Workshop on Research Issues in Data Engineering (RIDE2000)* (pp. 25–29).

Maekawa, M. (1985). A √n algorithm for mutual exclusion in decentralized systems. *ACM Transactions on Computer Systems*, *3*(2), 145-159.

Mat Deris, M., Evans, D. J., Saman, M. Y., & Noraziah, A. (2003). Binary vote assignment on a grid for efficient access of replicated data. *International Journal of Computer Mathematics*, *80*.

MAT DERIS, M. (2001). *Efficient access of replication data in distributed database systems*. PhD thesis, University Putra Malaysia.

Ozsu, M. T., & Valduriez, P. (1999). *Principles of distributed database system* (2nd ed.). Prentice Hall.

Reiher, P. P., Heidemann, J., Ratner, D., Skinner, G., & Popek, G. (1994, June). Resolving file conflicts in the Ficus file system. In *Proceedings of the Summer USENIX Conference* (pp. 183-195).

Walborn, S., & Chrysanthis, P. K. (1997). Pro-motion: Management of mobile transactions. In *Proceedings of the 11th ACM Symposium on Applied Computing*.

Wolfson, O., Jajodia, S., & Huang, Y. (1997). An adaptive data replication algorithm. *ACM Transactions on Database Systems*, *22*(2), 255-314.

Embedded Agents for Mobile Services

John F. Bradley
University College Dublin, Ireland

Conor Muldoon
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

Michael J. O'Grady
University College Dublin, Ireland

INTRODUCTION

A significant rise in the use of mobile computing technologies has been witnessed in recent years. Various interpretations of the mobile computing paradigm, for example, ubiquitous and pervasive computing (Weiser, 1991) and more recently, ambient intelligence (Aarts & Marzano, 2003)—have been the subject of much research. The vision of mobile computing is often held as one of "smart" devices operating seamlessly and dynamically, forming ad-hoc networks with other related devices, and presenting the user with a truly ubiquitous intelligent environment. This vision offers many similarities with the concept of distributed artificial intelligence where autonomous entities, known as agents, interact with one another forming ad-hoc alliances, and working both reactively and proactively to achieve individual and common objectives.

This article will focus on the current state of the art in the deployment of multi-agent systems on mobile devices and smart phones. A number of platforms will be described, along with some practical issues concerning the deployment of agents in mobile applications.

BACKGROUND

In the most general terms, an agent is one entity that acts, or has the authority to act, on behalf of another. In terms of information technology, an agent is a computational entity that acts on behalf of a human user, software entity, or another agent. Agents have a number of attributes that distinguish them from other software (Bradshaw, 1997; Etzioni & Weld, 1995; Franklin & Graesser, 1996; Wooldridge & Jennings, 1995):

- **Autonomy:** The ability to operate without the direct intervention from any entity, and possess control over their own actions and internal state.
- **Reactivity:** The ability to perceive their environment and react to changes in an appropriate fashion.
- **Proactivity:** The ability to exhibit goal-directed behavior by taking the initiative.
- **Inferential Capability:** The ability to make decisions based on current knowledge of self, environment, and general goals.
- **Social Ability:** The ability to collaborate and communicate with other entities.
- **Temporal Persistence:** The ability to have attributes like identity and internal state to continue over time.
- **Personality:** The ability to demonstrate the attributes of a believable character.
- **Mobility:** The ability to migrate self, either proactively or reactively, from one host device to another.
- **Adaptivity:** The ability to change based on experience.

An agent requires some space where it can exist and function, and this is provided for by an agent platform (AP). An AP comprises "the machine(s), operating system, agent support software,…agent management components…and agents" (FIPA, 2000, p. 6). The AP allows for agent creation, execution, and communication.

The majority of computer systems currently in operation use algorithms that are based on the concept of perfect information. The problem is that in the real world, businesses often require software functionality that is much more complex than this (Georgeff, Pell, Pollack, Tambe, & Wooldridge, 1999). Typically, computational entities within these systems should have an innate ability to deal with partial information and uncertainty within their environment. These types of systems are highly complex and are intractable using traditional approaches to software development. The rate at which business systems must change, due to market pressures and new information coming to light, requires software architectures and languages that more efficiently

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

manage the complexity that results from alterations being made to the code and the specifications.

Agent architectures, and in particular belief-desire-intention (BDI) (Rao & Geogeff, 1995) agent architectures, are specifically designed to deal with these types of issues and thus contain mechanisms for dealing with uncertainty and change. A problem with traditional systems is that they assume that they exist within a static or constant world that contains perfect information. The types of mobile systems that we are concerned with are dynamic and perhaps even chaotic, embedded with agents that have a partial view of the world and which are resource bounded.

Agents rarely exist in isolation, but usually form a coalition of agents in what is termed a multi-agent system (MAS). Though endowed with particular responsibilities, each individual agent collaborates with other agents to fulfill the objectives of the MAS. Fundamental to this collaboration is the existence of an Agent Communications Language (ACL), which is shared and understood by all agents. The necessity to support inter-agent communication has led to the development of an international ACL standard, which has been ratified by the Foundation for Intelligent Physical Agents (FIPA).

JAVA 2 MICRO EDITION (J2ME)

Most agent platforms developed for mobile devices have been written in the Java programming language—on mobile devices that usually means Java 2 Micro Edition (J2ME). This edition of Java contains a cut down API, a reduced footprint Java Virtual Machine, and a slightly different syntax (e.g., parameterized classes in Java 5). Java applications that contain dependencies on the idiosyncrasies of the different editions cannot be ported to a different range of devices without making alterations to the code. Their performance, however, is improved because the code is no longer developed to the lowest common denominator. Different algorithms and coding styles are now used for desktop machines and embedded devices rather than adopting comprised or overarching approaches that do not maximize the performance or maintainability of either.

A NUMBER OF AGENT PLATFORMS EXISTS FOR MOBILE DEVICES

3APL-M

3APL-M (Koch, 2005) is a platform that enables the fabrication of agents using the Artificial Autonomous Agents Programming Language (3APL) (Dastani, Riemsdijk, Dignum, & Meye, 2003) for internal knowledge representation. Its binary version is distributed in J2ME and J2SE compilations. 3APL provides programming constructs for implementing agents' beliefs, goals, basic capabilities, and a set of practical reasoning rules. The framework comprises an API that allows a Java application to call 3APL logic and deliberation structures.

Agent Factory Micro Edition

Agent Factory Micro Edition (AFME) (Muldoon, O'Hare, Collier, & O'Grady, 2006) is an agent platform developed for the construction of lightweight intelligent agents for cellular digital mobile phones and other compatible mobile devices. AFME is broadly based on Agent Factory (Collier, 2001), a pre-existing J2SE framework for the fabrication and deployment of agents. AFME differs from the original version of the system in that it has been designed to operate on top of the Constrained Limited Device Configuration (CLDC) Java platform augmented with the Mobile Information Device Profile (MIDP). CLDC and MIDP form a subset of the J2ME platform specifications. Though sharing the same broad objectives of the other projects mentioned in this section, AFME differs in a number of ways. With a jar size of 85k, it is probably the smallest footprint FIPA-compliant deliberative agent platform in the world. The platform supports the development of a type of software agent that is: autonomous, situated, socially able, intentional, rational, and mobile. An agent-oriented programming language and interpreter facilitate the expression of an agent's behavior through the formal notions of belief and commitment. This approach is consistent with a BDI agent model.

LEAP

Probably the most widely known agent platform for resource-constrained devices is the Light Extensible Agent Platform (LEAP) (Berger, Rusitschka, Toropov, Watzke, & Schichte, 2002). LEAP is a FIPA-compliant agent platform developed to be capable of operating on both fixed and mobile devices with various operating systems in wired or wireless networks. Since version 3.0, LEAP extends the Java Agent DEvelopment Framework (JADE) (Bellifemine, Caire, Poggi, & Rimassa, 2003) by using a set of profiles that allow it to be configured for various Java Virtual Machines (JVMs). The architecture of the platform is modular and contains components for managing the lifecycle of the agents and controlling the array of communication protocols. The platform is split into several agent containers, one for every device used. These containers are responsible for passing messages between agents and choosing the appropriate communication protocol. One of these containers, known as the main container, includes agents that fulfill the white and yellow pages services as necessitated by the FIPA specification.

MAE

The MAE (mobile agent environment) (Mihailescu, Binder, & Kendall, 2002) agent platform has been designed to be independent of device and language implementations. To accomplish this, the platform is divided into a reference API specification, reference implementation, and non-standard implementation additions. The reference API specification is not dependent on programming language or hardware, and it contains the core platform components. The Reference Implementation contains all the device-dependent code required by the reference API specification. The third part, non-standard implementation additions, is used for application-specific components.

While this approach gives a high degree of platform independence, unless it is being deployed in an environment of homogeneous devices, it means a lot of work as each platform may require its own implementation.

MicroFIPA-OS

MicroFIPA-OS is an agent toolkit based on the standard FIPA-OS but optimized for resource-constrained mobile devices (Tarkoma & Laukkanen, 2002). It targets the personal Java platform and thus operates on personal data assistants. The system can run in minimal mode whereby agents do not use task and conversation managers. Yellow and white page services are provided in compliance with the FIPA specification. The platform is entirely embedded, however it is recommended that only one agent operate on low-specification devices.

NON-EMBEDDED AGENTS FOR MOBILE SERVICES

There are other types of agent platforms suitable for mobile services that do not embed the agents in the mobile device:

- platforms that use the mobile device as just an interface while the agents are executed on more capable hosts, for example MobiAgent; and
- platforms that do part of the execution on the mobile device, while simultaneously executing the remainder the task on other hosts such as KSACI (Hübner, 2000a, 2000b).

MobiAgent

A MobiAgent (Mahmoud, 2001) platform comprises a handheld mobile wireless device and an agent gateway, which are networked and communicate through hypertext transfer protocol (HTTP). The agent gateway executes the agent and its associated apparatus. The user interacts with the agent through an interface on the mobile device, which connects to the agent gateway and configures the agent. After the agent carries out a task, it reports back through the interface. This approach requires the minimum amount of processing and memory resources on the mobile device, but it makes the connectivity essential.

KSACI

Simple agent communication infrastructure (SACI) is a framework for creating agents that communicate using the Knowledge, Query, and Manipulation Language (KQML) (Finin, 1997). Each SACI agent has a mailbox to communicate with other agents. Infrastructure support is provided for white and yellow pages, but the platform is not FIPA compliant. KSACI is a smaller version of SACI suitable for running on the kVM (Albuquerque, Hübner, de Paula, Sichman, & Ramalho, 2001). The platform is not entirely situated on the constrained device and only supports the running of a single agent, which communicates via HTTP with a proxy running on a desktop machine.

DISCUSSION

A mobile computing environment is typified by resource constraints. Issues like processing power, memory, battery life, connectivity, and input/ouput (I/O) all require careful consideration.

It is often reported that intelligent agent platforms are unsuited for mobile applications because of their excessive computational overhead. This problem is usually due to particular agent platform implementations rather than an innate problem with the agent paradigm itself. Improving the efficiency of the reasoning algorithms within these systems can often lead to significant gains in efficiency. Additionally, the programming style adopted by the developer can have a considerable impact on performance. Developing in a style that conforms to the Law of Demeter (Lieberherr, Holland, & Riel, 1988) can reduce the footprint of the software by minimizing duplicated code while also improving maintainability in that internal implementation details of the object model are hidden. Further performance gains may be obtained through the use of autonomic procedures. An example of such a procedure, termed Collaborative Agent Tuning, may be found in Muldoon, O'Hare, and O'Grady (2005). Tuning enables agents collectively to alter their response times and computational overhead so as to maximize system performance.

The communication infrastructure is another fundamental resource that must be managed astutely when developing multi-agent systems. It is particularly important when working with lightweight devices that have limited battery power since sending messages consumes significantly more battery resources than normal processing. Mobile devices often have limited bandwidth and must make intelligent decisions as to what information to download and when to download it.

Additionally there is the issue of human-agent communication. Consideration must be made for the I/O capabilities of the devices. Most would have some form of keyboard input in the form of a touch screen or keypad. How the agent would convey information would be a bigger modality issue—is there a screen, does it allow for graphics or just text, how big is the screen, and how much of it is available to the agent?

FUTURE TRENDS

In the future, agents will emerge that are endowed with autonomy, mobility, and human-computer interaction facilities (Bradley, Duffy, O'Hare, Martin, & Schön, 2004). Such agents will opportunistically migrate, based on their tasks at hand, to different platforms (each offering varying capabilities and prospects), which would usually be for the benefit of an associated user. The presence of the agent moving through cyberspace as the user moves through physical space allows the associated user to be contactable at anytime through the agent.

A clear application of such nomadic agents is that of an autonomous "intelligent" digital assistant that is independent of any one physical device. These entities will effectively give any user his or her own personal assistant that will help with the information overload in daily life, assisting with personal communications and offer a generic interface to any number of devices. These devices will have the ability to react to the current needs of their user, and beyond this, grow and learn to anticipate future needs and requirements. Perhaps our vision can be best summed up by Luc Steels' metaphor for what the robots of the future will be like:

> *[It] is related to the age-old mythological concept of angels. Almost every culture has imagined persistent beings which help humans through their life. These beings are ascribed cognitive powers, often beyond those of humans, and are supposed to be able to perceive and act in the real world by materialising themselves in a bodily form at will.* (Steels, 1999)

He goes on to detail how angels may "project the idea of someone protecting you, preventing you from making bad decisions or actions, empowering you, and defending you in places of influence."

CONCLUSION

Agents encapsulate a number of features that make them an attractive and viable option for realizing mobile services. At a basic level, their autonomous nature, ability to react to external events, as well as an inherent capability to be proactive in fulfilling their objectives make them particularly suitable for operating in complex and dynamic environments. Should an agent be endowed with a mobility capability, its ability to adapt and respond to unexpected events is further enhanced. However, there are a few negative aspects to using agents. These systems can be more complex and require more device resources than the equivalent application-specific programs. Having no native support, agents require their own agent platforms for creation, execution, and communication. These problems will be reduced with advancements in mobile computing technologies, however in order to optimize system performance, agents will still have to manage their resources in a prudent and intelligent manner.

REFERENCES

Aarts, E., & Marzano, S. (Eds.). (2003). *The new everyday: Views on ambient intelligence.* Rotterdam, The Netherlands: 010.

Albuquerque, R. L., Hübner, J. F., de Paula, G. E., Sichman, J. S., & Ramalho, G. L. (2001, August 1-3). KSACI: A handheld device infrastructure for agents communication. *Pre-proceedings of the 8th International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, Seattle, WA.

Bellifemine, F., Caire, G., Poggi, A., & Rimassa, G. (2003, September). *JADE.* White Paper.

Berger, M., Rusitschka, S., Toropov, D., Watzke, M., & Schichte, M. (2002). Porting distributed agent-middleware to small mobile devices. *Proceedings of the Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices held in conjunction with the Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Bologna, Italy.

Bradley, J. F., Duffy, B. R., O'Hare, G. M. P., Martin, A. N., & Schön, B. (2004, September 7-8). Virtual personal assistants in a pervasive computing world. *Proceedings of IEEE Systems, Man and Cybernetics, the UK-RI 3rd Workshop on Intelligent Cybernetic Systems (ICS'04)*, Derry, Northern Ireland.

Bradshaw, J. M. (1997). An introduction to software agents. In J. M. Bradshaw (Ed.), *Software agents* (pp. 3-46). Boston: MIT Press.

Collier, R. W. (2001, March). *Agent factory: A framework for the engineering of agent-oriented applications.* PhD thesis, Department of Computer Science, University College Dublin, National University of Ireland.

Dastani, M., Riemsdijk, B., Dignum, F., & Meye, J. J. (2003). A programming language for cognitive agents: Goal directed 3APL. *Proceedings of the 1st Workshop on Programming Multiagent Systems: Languages, Frameworks, Techniques, and Tools* (ProMAS), Melbourne.

Etzioni, O., & Weld, D. S. (1995). Intelligent agents on the Internet: Fact, fiction, and forecast. *IEEE Expert, 10*(4), 44-49.

Finin, T., & Labrou, Y. (1997). KQML as an agent communication language. In J. M. Bradshaw (Ed.), *Software agents* (pp. 291-316). Boston: The MIT Press.

FIPA (Foundation for Intelligent Physical Agents). (2000). *FIPA agent management specification.* Retrieved from http://www.fipa.org

Franklin, S., & Graesser, A. (1996). Is it an agent or just a program? A taxonomy for autonomous agents. *Proceedings of the 3rd International Workshop on Agent Theories, Architectures, and Languages.* New York: Springer-Verlag.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. (1999). The belief-desire-intention model of agency. *Proceedings of the 5th International Workshop on Intelligent Agents V: Agent Theories, Architectures, and Languages (ATAL-98)*, Paris, France.

Hübner, J. F., & Sichman, J. S. (2000a). SACI: Uma ferramenta para implementação e monitoração da comunicação entre agents. *Proceedings of IBERAMIA.*

Hübner, J. F., & Sichman, J. S. (2000b). *SACI programming guide.*

Koch, F. (2005, July 25-29). 3APL-M platform for deliberative agents in mobile devices. *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS"05)* (pp. 153-154), The Netherlands. New York: ACM Press.

Lieberherr, K. J., Holland, I., & Riel, A. J. (1988). Object-oriented programming: An objective sense of style. Object oriented programming systems, languages and applications conference. *SIGPLAN Notices (Special Issue),* (11), 323-334.

Mahmoud, Q. H. (2001). MobiAgent: An agent-based approach to wireless information systems. *Proceedings of the 3rd International Bi-Conference Workshop on Agent-Oriented Information Systems,* Montreal, Canada.

Mihailescu, P., Binder, W., & Kendall, E. (2002). MAE: A mobile agent platform for building wireless m-commerce applications. *Proceedings of the 8th ECOOP Workshop on Mobile Object Systems: Agent Applications and New Frontiers,* Málaga, Spain.

Muldoon, C., O'Hare, G. M. P., Collier, R. W., & O'Grady, M. J. (2006, May 28-31). Agent factory micro edition: A framework for ambient applications. *Proceedings of Intelligent Agents in Computing Systems, a Workshop of the International Conference on Computational Science (ICCS 2006)*, Reading.

Muldoon, C., O'Hare, G. M. P., & O'Grady, M. J. (2005). Collaborative agent tuning. *Proceedings of the 6th International Workshop on Engineering Societies in the Agents' World (ESAW 2005)*, Kusadasi, Turkey.

Rao, A. S., & Georgeff, M. P. (1995, June). BDI agents: From theory to practice. *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS'95)* (pp. 312-319), San Francisco.

Steels, L. (1999). *Digital angels.* Retrieved from http://arti.vub.ac.be/steels/sued-deutsche.pdf

Tarkoma, S., & Laukkanen, M. (2002). Supporting software agents on small devices. *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Bologna, Italy.

Weiser, M. (1991). The computer for the twenty-first century. *Scientific American,* (September), 94-100.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review, 10*(2), 115-152.

KEY TERMS

Agent: A computational entity that acts or has the authority to act on behalf of a human user, software entity, or another agent.

Agent Communication Language: A formal language used for communication between agents.

Agent Platform: Provides the necessary infrastructure on which an agent operates.

Ambient Intelligence: Computing and networking technology that is unobtrusively embedded in the environment.

Embedded Agent: An agent that is contained wholly, along with its platform, on a particular device.

Mobile Service: One of several services provided through devices in a mobile computing environment (i.e., mobile phones, personal data assistants, wearable computers, etc.).

Multi-Agent System: A system comprising several agents—on the same platform or across multiple platforms—with a common goal.

Pervasive Computing: Computing involving computers, usually mobile devices, in all aspects of daily life.

Ubiquitous Computing: Computing in which the computers are embedded in everyday objects and all computing is done in the background.

Enabling Mobile Chat Using Bluetooth

Ádrian Lívio Vasconcelos Guedes
Federal University of Campina Grande, Brazil

Jerônimo Silva Rocha
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

Mobile chat applications can be seen as an alternative and effective way of communicating for people without the need of using the mobile telephony system. Based on the new generation of cellular phones with support for communication technologies, such as Bluetooth and Wi-Fi, it is possible to develop applications to enable mobile chats. Such applications can provide mechanisms to discover and communicate with other devices in a shorter range, but with low or no communication costs.

This article introduces Let's Talk, a mobile chat and relationship application. It allows a Symbian OS Series 60 mobile phone user to create a profile and share it with other users. Also, it is possible to invite other users to a chat in a session. The profile sharing and the chat communication data are transferred over a Bluetooth connection. After creating your profile, a user can search for other profiles in the range of the Bluetooth connection and make your profile available to other users.

In this article, we discuss design and implementation issues related to the application development using a Symbian-based cellular phone and the C++ programming language. The remainder of this article is organized as follows. We first present the technologies used to develop the application: Symbian OS, the Series 60 platform, the Bluetooth wireless technology, and the Cobain Framework. We then present the Let's Talk software and the use of the technologies presented in the Background section. Possible improvements for the application and trends related to the theme are then offered, followed by final remarks in the Conclusion section.

BACKGROUND

Symbian OS

The Symbian OS (http://www.symbian.com/) is an operating system designed for mobile devices; it is an industry standard, used in smart phones of many manufacturers, such as Nokia, Siemens, Motorola, Sansung, and others.

Symbian is optimized for mobile devices that have low memory and processing power, with low runtime memory requirements. It is designed to optimize the device performance and the battery life. It is a multi-tasking operating system, allowing many applications to run concurrently. To reduce recourse consumption, Symbian provides multi-thread support to the programmer through the concept of active objects, which are a lightweight alternative to threads.

The Symbian OS development model is based on an object-oriented architecture using the C++ programming language with optimized memory management for embedded software.

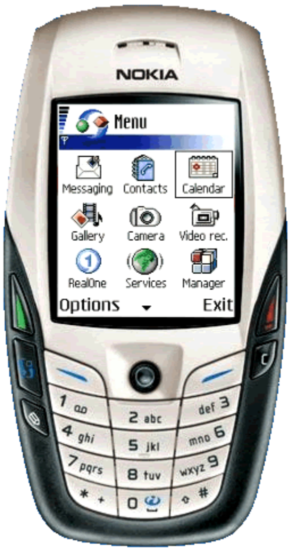
Series 60

The Series 60 platform (http://www.s60.com/) was developed by Nokia, but it is also licensed to other manufacturers. It was built over the Symbian Operating System, providing a configurable graphical user interface library and a set of applications and other general-purpose implementations.

The set of applications includes personal information management (PIM) and multimedia applications, such as calendars, contacts, text and multimedia messaging (SMS,

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Series 60 user interface



MMS), e-mail, browsing using WAP or others, and so forth.

Some of the main features of Series 60 are the large color screen with a minimum specification of 172 by 208 pixels, and at least 4,096 colors (64K colors in Series 60 2.x) and many interaction models, such as two soft keys, five-way navigator, and other dedicated keys (Edwards, Barker, & Staff of EMCC Software, 2004). The Series 60 User interface is illustrated in Figure 1.

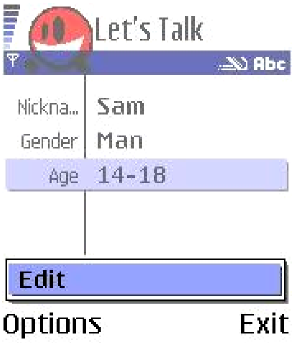
Bluetooth

Bluetooth (IEEE 802.15.1) is a wireless specification for personal area networks (PANs). It provides a way to connect and exchange information between devices such as personal digital assistants (PDAs), mobile phones, laptops, PCs, printers, and digital cameras via a secure, low-cost, globally available short-range radio frequency (http://www.bluetooth.com/). Bluetooth is available in most Series 60 devices providing connectivity to these devices.

The Cobain Framework

Cobain is an API (application programming interface) that permits the development of Bluetooth applications, simplifying the development process for this kind of application in the Symbian OS (Dahlbom & Kokkola, 2004). It consists of a lightweight ad-hoc networking framework, providing a Unix-like API socket, hiding details of implementation, such as Active Objects handling.

Figure. 2 Profile form



Nickna... Sam
Gender Man
Age 14-18

LET'S TALK

Let's Talk is a chat application for Symbian OS Series 60 mobile phones, allowing users to contact another people and establish a conversation. The profile sharing mechanism allows a user to create a profile and share it with other users. The profile contains personal information, such as name, age, and gender. This information is made available to other Let's Talk users that can invite this user to chat after viewing his/her profile. This feature enables the establishment of relationships between users based on the level of interest in their profiles.

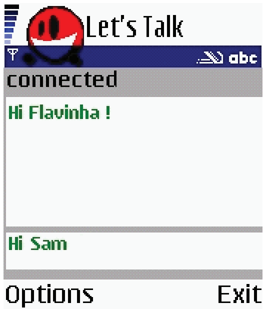
The application may be running in two modes: waiting or searching. The searching application can discover all devices running the application in the waiting mode and allows the user to request the profile of any discovered user. The profile request contains the searching user profile data, which will be evaluated by the waiting user.

The waiting mode informs the user that there is an incoming profile request and shows the profile of the requesting user in a form. In response to this request, the user sends the waiting user profile data to the searching application, which shows this profile to the searching user.

In each device, Let's Talk creates a form that contains the profile of the other user using the data sent via Bluetooth. In the Series 60 platform, the forms provide a way for the user quickly and easily to enter or edit many items of data in the application. The form could also be used in *view mode* to display information about the user profile. If a form has a view mode, like the profile creation form, the form focus appears as a solid block as illustrated in Figure 2. Switching to edit mode is achieved by selection of Edit from the Options menu (Edwards et al., 2004).

The profile sharing and the chat communication data are transferred over a Bluetooth connection. The connection is established using the Cobain API that is responsible for: (1) discovering devices available for Bluetooth connection, (2) discovering services available at the selected device, (3) connecting to the given service, (4) sending and receiving

Figure 3. Chat user interface



Hi Flavinha !

Hi Sam

the chat and profile data to and from the remote service, (5) and closing the connection (Dahlbom & Kokkola, 2004). The chat user interface is depicted in Figure 3.

The application can be used in several places like restaurants, shopping centers, subways, coffee shops, and other places with great people concentration. For these various scenarios the application behavior is similar, but improvements could be performed for specific contexts. For example, the profile of chat applications for restaurants could include preferences related to gastronomy.

FUTURE TRENDS

Application Improvements

In a future version of the application, some new functionalities may be implemented, such as search by profile information, context-sensitive profile, and automatic profile comparison.

- **Search by Profile Information:** The application can provide a search field for specific profile information. A user can search for all users that have a specific age, for example.
- **Context-Sensitive Profile:** The profile form fields may change depending on the application scenario. In a professional ambient such as an academic lab, the user profile may contain information about his/her skills in a specific area. Another user with a problem in a specific knowledge area can find someone to help him/her.
- **Automatic Profile Comparison:** The application may automatically share the profile and compare the contents. In a party, a user may be advertised that there is another user in this party that likes the same music style as him/her.
- **Server Connection:** Using a Bluetooth access point, the application may have access to a central service that can provide functionalities like profile database access, or improve the range of application over than a Bluetooth connection.

Future Trends in Mobile Chat

In the context of mobile chat, future trends include the usage of new technologies related to multimedia, content, and connection mechanisms. Considering connection mechanisms, Bluetooth architecture could work together with wired infrastructure to enable long-range chat.

A multimedia-enabled mobile chat application can provide creation, exchange, and presentation of videos, sounds, and images during a chat session. Video and audio chats could also be possible, without using the telephony infrastructure. For these applications, though, mechanisms of lower battery consumption are fundamental.

Access to content and services of wired infrastructures during chat sessions is also an interesting feature to be considered. A chat user could obtain content related to his/her conversation and services to support it, like dictionaries and translators.

CONCLUSION

Let's Talk is an interesting application for personal relationships, providing costless, wireless communication. The profile sharing mechanism enables users to choose persons to chat with based on profile descriptions.

With the improvements proposed earlier in this article, the application can become a powerful tool enabling the user to find and communicate with persons of interest.

The enhancement of new technologies enables the applications to have access to Web and multimedia contents. These contents also may be used in mobile chat applications like Let's Talk.

REFERENCES

Dahlbom, M., & Kokkola, M. (2004). *Cobain architectural specification.* Retrieved from http://irssibot.777-team.org/cobain/index.html

Edwards, L., Barker, R., & Staff of EMCC Software. (2004). *Developing Series 60 applications: A guide for Symbian OS C++ developers.* Boston: Addison-Wesley.

KEY TERMS

API: Application programming interface.

IEEE: Institute of Electrical and Electronics Engineers.

MMS: Multimedia message service.

PAN: Personal area network.

PDA: Personal digital assistant.

PIM: Personal information management.

SMS: Short message service.

WAP: Wireless application protocol.

Enabling Mobility in IPv6 Networks

Saaidal Razalli Bin Azzuhri
Malaysia University of Science and Technology, Malaysia

K. Daniel Wong
Malaysia University of Science and Technology, Malaysia

INTRODUCTION

With the explosive growth in Internet usage over the last decade, the need for a larger address space is unavoidable, since all the addresses in IPv4 are nearly fully occupied. IPv6 (Deering & Hinden, 1998), with 128-bit addresses compared to IPv4 with 32-bit addresses and other advantages (like auto-configuration and IP mobility), can overcome many of the problems that IPv4 had before.

One of the requirements for the modern Internet is IP mobility support. In IPv4, a special router is needed to act as a foreign agent in the visited/foreign network and the need of a network element in the home network known as a home agent for a mobile host. IPv6 does away with the need for the foreign agent and operates in any location without any special support from a local router. Route optimization is inherent in IPv6, and this feature eliminates the triangle-routing (routing through the home agent) problem that exists in IPv4. IPv6 enjoys many network optimizations that are already built in within IPv6.

IP MOBILITY

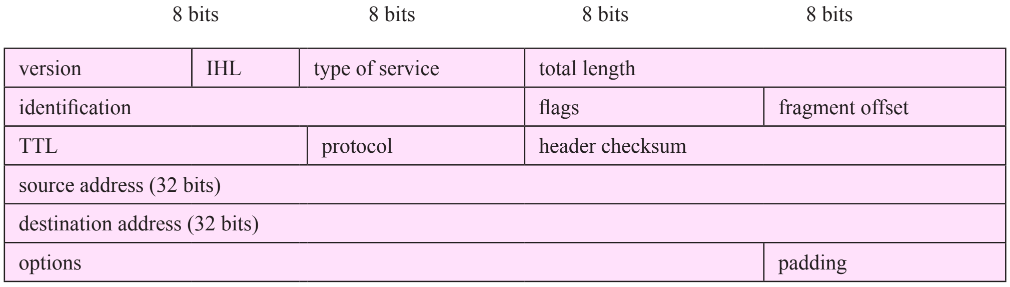
IP mobility can be defined as referring to situations where there is a change in a node's IP address due to a change of its attachment point within the Internet topology (Soliman, 2004). This change may be caused by physical moment, such as someone moving her computer from one room to another or someone sitting in a moving vehicle that traverses different links. IP mobility can also occur due to change in the topology, which causes a node to change its address. Mobile IPv6 is a suite of protocols for IPv6 nodes to handle IP mobility.

Mobile IPv6 allows an IPv6 host to leave its home subnet while transparently maintaining all its connections and remaining reachable to the rest of the Internet. The use of IP in wireless technologies, such as local area networks (LANs; e.g., IEEE 802.11a, b, and g) to wide area networks (WANs; e.g., 3G), makes mobility in wireless devices an interesting research field. The popularity of wireless technologies allows users (hosts) to move freely within large geographical areas, but requires good support for mobility. Mobile IPv6 is the more prominent solution for mobility for IP wireless devices (Samad & Ishak, 2004). We will first review some relevant features of IPv6 before explaining how Mobile IPv6 works.

RELEVANT FEATURES OF IPV6

IPv6 specification was already defined in RFC 2460 (Deering & Hinden, 1998). Figures 1 and 2 show the header comparison between IPv4 and IPv6 headers. These include the header format, extension headers, and their processing fields. As can be seen in the figures, only the IPv6 header contains the minimum amounts of information necessary for IPv6 hosts to communicates with each other. The IPv6 packet header is much simpler than the IPv4 one. It is now a fixed size with no optional fields. Options in IPv4 are replaced by an IPv6

Figure 1. IPv4 header



8 bits 8 bits 8 bits 8 bits

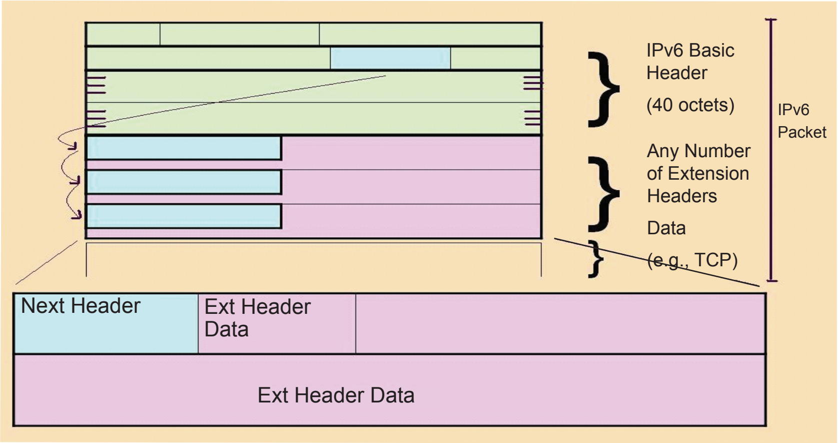
version	IHL	type of service	total length	
identification			flags	fragment offset
TTL	protocol		header checksum	
source address (32 bits)				
destination address (32 bits)				
options				padding

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. IPv6 header

8 bits	8 bits	8 bits	8 bits
version	traffic class	flow label	
payload length		next header	hop limit
source address (128 bits)			
destination address (128 bits)			

Figure 3. Extension header



extension header (will be explained later), which includes additional parameters for hosts or routers to receive IPv6 packets. An extension header in IPv6 may contain one or more extension headers when necessary for the processing of such a packet (Deering & Hinden, 1998).

The IPv6 header removes some of the fields that were previously included in the IPv4, and added new fields. It has been slimmed down to the necessary minimum header compression, which now has 8 fields, compare to the previous 13 fields in IPv4. The following sections explain further some of these design choices.

Extension Header

Mobile IPv6 has an optional header called *extension header.* IPv6 extension headers are defined to encode certain options that are needed for processing of the IPv6 packet and its subsequent packets. Encoding options must minimize the amount of time needed in order to classify the header and forward the packet on the correct route. The benefits of extension headers can be best explained when comparing them to option fields in IPv4 headers. Consider the router receiving an IPv4 packet including one or more options. The router would first determine that the packet is carrying IP options. The next step is the router must parse or classify the IP header to find out which options require processing by the router itself, as opposed to processing by the ultimate receiver of the packet. The process of parsing this header and its options takes some time and can reduce efficiency.

Routing Header

IPv6 defines a fixed size 40-bytes header and extension header for additional options. The routing header includes addresses of nodes that must be in the path taken by a packet on its way to its ultimate destination. Thus, the routing header is a form of source routing and can be used to make sure the packet goes to certain nodes/addresses on its way to its ultimate destination. It also allows routing to certain special-purpose routers for special reasons (e.g., mobility support).

Hop-by-Hop Options Header

As the name implies, this header includes options that need to be processed by every node (routers) along the packet's delivery path. It specifies delivery parameters at each hop on the way to the destination. Some of the fields in this type of header are used to alert a router to things like multicast listener discovery, that is, that this packet is part of a multicast and requires special processing.

Destination Options Header

The destination option is used to specify a process that needs to be performed by the destination node, whose address is the destination address in the IPv6 header. It is useful for Mobile IPv6 as the destination options header used to exchange the registration messages between mobile nodes and the home agent. Delivery parameters are either intermediate hops or the final destination, similar to the hop-by-hop options. The difference between destination and hop-by-hop options is that the former is processed by nodes that the packets are destined to, while the latter is processed by every node along the network path until the last receiver.

Authentication Header

The authentication header is a mechanism to protect a packet's integrity following the establishment of a security association. It is used to provide data authentication and integrity checking information, but not encryption. In addition, the authentication header protects against replay attacks by including a sequence number field, which is incremented each time the packets are sent. It is mechanism of security in IPv6, not fully bullet-proof, but it provides first level of data security (Faccin & Le, 2003).

Fragment Header

Fragment header is used similarly as in IPv4. It indicates that this packet is part of a fragmented stream, but fragmentation is only allowed on the part of the sender. Routers are not allowed to fragment payloads, which makes for better quality-of-service overall. In IPv6, only the sending host can perform this function.

Tunneling

Tunneling can be defined as a process whose node (host or router) encapsulates an IPv6 packet in another IPv6 header, which can be two or more packets (if encapsulation is done more than once) (Jeong, Park, & Kim, 2004). There are several terms associated with tunneling:

- **Tunnel Entry Point:** Originating tunnel node.
- **Tunnel Exit Point:** Terminating tunnel mode.

The tunnel will act as a virtual point-to-point link when seen by the original IPv6 header, starting at the tunnel entry point and ending at the tunnel exit point. The header of the new IPv6 packet is shown in Figure 4. Note that the tunnel exit point decapsulates the packet, and the decapsulated packet will be sent to the host that its destined to.

Tunneled IPv6 Packet

When an IPv6 packet is tunneled inside a new IPv6 packet, the router along the tunnel point will only recognize the outer header, which contains a new source and destination address. Tunneling is a very important mechanism in mobility, which will be discussed later. When packets arrive at a tunnel exit point, the outer header is thrown away by the tunnel exit router, and the original header will be processed (decapsulation process). Mobile IPv6 makes use of tunnels where the source address of tunneled packets will be the home agent address and the destination address will be the mobile node (MN) care-of-address (COA).

Router Discovery

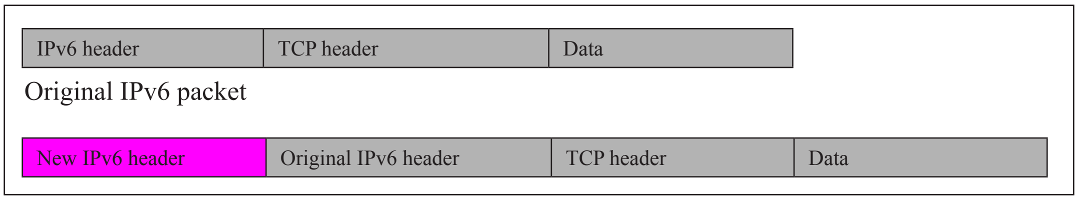
Router Solicitations Mechanism

A router solicitation is used by hosts to discover one or more default routers neighboring the solicitations host. It is sent to an all-routers multicast address (link-local multicast address). This link-local multiple address is hardcoded in all router implementations.

Router Advertisement Mechanism

A router advertisement can be described as a response to router solicitations. When a router is solicited, the advertisement is sent to a unicast address of soliciting node. By contrast,

Figure 4. IPv6 tunneling



an unsolicited router advertisement is sent to 'all-nodes multicast addresses'.

Stateless Address Autoconfiguration

Using the prefix information obtained from a router advertisement (RA), nodes can append an EUI-64 bit interface identifier to the advertised prefix to form a unique IPv6 address. Thus, it can obtain a global address in a stateless way. The term stateless refers to the fact that the address is configured without the need for keeping a record for such address allocation in any node (no state), except for the node that assigned the address to one of its interfaces. This eliminates the need for a stateful server like a DHCP server that keeps track of addresses allocated on the link.

HOW IPV6 MOBILITY WORKS

Mobile IPv6 allows IPv6 nodes to be mobile. It also allows nodes to be reachable and maintain ongoing connections while changing their location within the network topology (Silva, Camilo, Costa, Matos, & Boavida, 2004). Connection maintenance is done by the IP layer using Mobile IPv6 messages, options, and processes that ensure correct delivery of data regardless of the mobile node's location. This operation is transparent to upper layers, in order to maintain sessions as the mobile node changes its location. It is important to understand the components of Mobile IPv6 first before going on to details on mobility operation.

Since the Internet protocol itself does not address node mobility, modification is needed to enable nodes to continue to receive packets when they change their points of attachment to the Internet. This raises the need for a mobile Internet protocol. The mobility support protocol for IPv6 developed by IETF is known as Mobile IPv6 (RFC3775) (Johnson, Perkins, & Arkko, 2004). Just like the relationship between IPv6 and IPv4, Mobile IPv6 evolved from its counterpart known as Mobile IPv4 (or just Mobile IP for short). However, several new features in IPv6 make it more accommodating to mobility than IPv4.

Mobile IPv6 Components

Mobile Node (MN)

Mobile node is an IPv6 node that can change links, and therefore addresses, while maintaining reachability using its home address. A mobile node has awareness of its home address and the global address for the link to which it is attached (known as the care-of address), and indicates its home address/care-of address mapping to the home agent and Mobile IPv6-capable nodes with which it is communicating.

Correspondent Node (CN)

Correspondent node is an IPv6 node that communicates with a mobile node. A correspondent node does not have to be Mobile IPv6-capable. However, if the correspondent node is Mobile IPv6-capable, it can also be a mobile node that is away from home.

Home Address

An address is assigned to the mobile node when it is attached at the home link and through which the mobile node is always reachable, regardless of its location on an IPv6 network. If the mobile node is attached to the home link, Mobile IPv6 processes are not used, and communication occurs normally. If the mobile node is away from home (not attached to the home link), packets addressed to the mobile node's home address are intercepted by the home agent and tunneled to the mobile node's current location on an IPv6 network. Because the mobile node is always assigned the home address, it is always logically connected to the home link, even if it is physically somewhere else.

Home Link (HL)

This is a link to which a home address prefix is assigned, from which the mobile node obtains its home address. The home agent resides on the home link.

Foreign Link (FL)

This is a link that is not the mobile node's home link, but one that is visited by mobile node.

Home Agent (HA)

A home agent is a router on the home link that maintains registrations of mobile nodes that are away from home and the different addresses that they are currently using. If the mobile node is away from home, it registers its current address with the home agent, which tunnels data sent to the mobile node's home address to the mobile node's current address on an IPv6 network and forwards tunneled data sent by the mobile node.

Care-of-Address (COA)

A COA is an address used by a mobile node while it is attached to a foreign link. For stateless address configuration, the care-of address is a combination of the foreign subnet prefix and an interface ID determined by the mobile node. The association of a home address with a care-of address for a mobile node is known as a *binding.* Correspondent nodes and home agents keep information on bindings in a *binding cache.*

Overall components of Mobile IPv6 and their placement in a typical network can be seen in Figure 5. Please note that the mobile node itself can be in the home link, in which case it will be using its home IPv6 address and normal IPv6 packet forwarding. Mobile IPv6 allows an IPv6 hosts to leave its home, while transparently maintaining all its connections and remaining unreachable to the rest of the Internet.

Mobile IPv6 Procedures

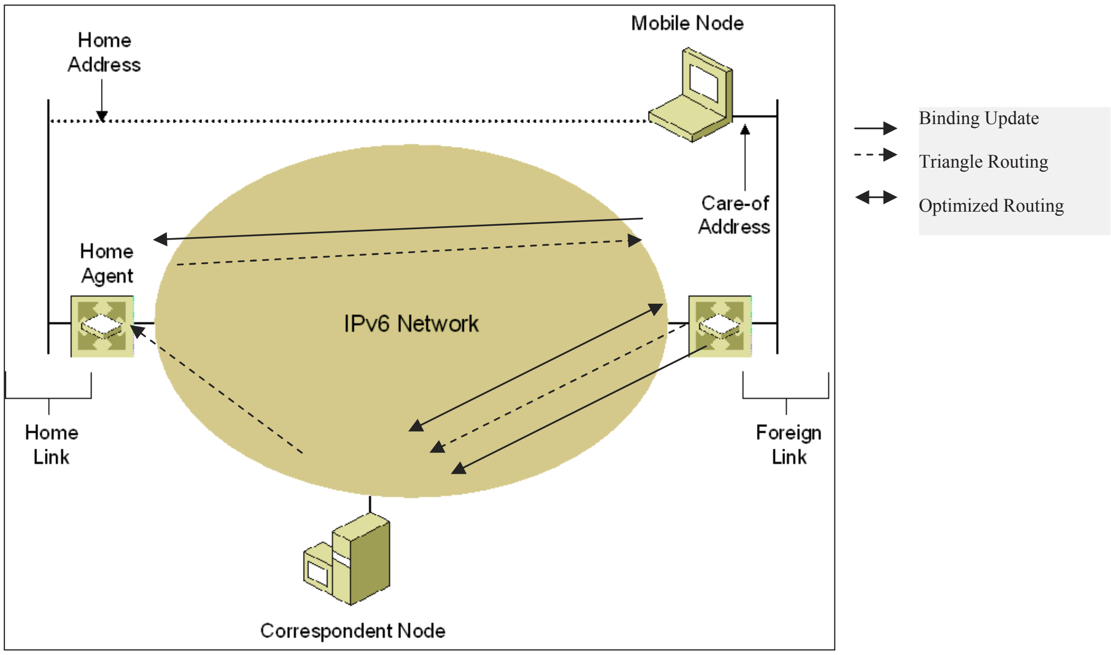
Figure 5 shows the Mobile IPv6 architecture. A mobile node (MN) first needs to determine whether it is currently connected to its home link or a foreign link. If it detects it has moved to a foreign link, it will obtain a care-of-address; it also reports its COA to the correspondent node (CN). These two procedures are called *binding update,* and once acknowledged (same as binding update, but in reverse direction, known as 'binding acknowledgement'), binding update with the correspondent node will be known as *route optimization.* Once the CN knows the mobile node's COA, it will be able to send further packets directly to mobile node's COA, without going through the triangle route via the mobile node's home agent as shown in Figure 5. The above brief overview of Mobile IPv6 operation contains three key components of the protocol, namely, *router discovery, address notification,* and *packet routing,* which will be further illustrated in the following sections.

Mobile IPv6 Router Discovery

Router discovery has three main functions for a mobile node. Firstly, router discovery determines whether the mobile node is currently connected to its home link or a foreign link. Two types of messages, *router advertisements* and *router solicitation,* are involved. Router advertisements are used by the routers and home agents to announce their capabilities to mobile nodes. Specifically, a router advertisement is periodically transmitted as broadcast on each attached network where the node is configured to perform as a home agent or a Mobile IPv6 router, or both. Router solicitations are sent by mobile nodes that do not have the patience to wait for the next periodic transmission of a router advertisement. So the only purpose of a router solicitation is to force any routers or home agents on the network to immediately transmit a router advertisement. This is useful when the frequency at which routers and home agents are transmitting is too low compared with the moving frequency of the mobile node.

Secondly, router discovery can detect whether the mobile node has moved from one network to another. The mobile node may perform location and movement detection by examining the network prefixes contained in a received advertisement. If any of these prefixes matches the network-prefix of the mobile node *home address,* then the mobile node is connected to its home link; otherwise, if none of the prefixes matches the network prefixes of the mobile node's home address, then the mobile node decides that it is connected to a foreign link.

Figure 5. Components of Mobile IPv6



Thirdly, router discovery helps the mobile node obtain a COA in the foreign link. There are two methods by which a mobile node can acquire a COA in IPv6. The first method is *stateful address autoconfiguration,* in which the mobile node simply asks a server for an address and uses it as a care-of-address. The second method of acquiring a COA is *stateless address autoconfiguration,* in which a mobile node automatically forms a COA by concatenating a network prefix with an interface token. The first method is very similar to DHCP in IPv4 (the name is converted to DHCPv6), while the second is new to IPv6. The choice of specific method depends on the information contained in the received router advertisements.

Mobile IPv6 Address Notification

The Mobile IPv6 address notification is the process by which a mobile node informs both its home agent and various correspondent nodes of its current COA (Chirovolu, Agrawal, & Vandenhoute, 1999). The home agent uses the COA as the tunnel destination to forward a packet to the mobile node when it is away from home. The correspondent node uses the COA to route packets directly to the mobile node, without going through the mobile node's home agent. Thus, Mobile IPv6 has a built-in route optimization.

The messages used for notification include *binding request, binding update,* and *binding acknowledgment* (Qi, 2001). A binding request sent by a correspondent node to a mobile node can also initiate the binding update from it. Otherwise, the mobile node can also initiate the binding update by itself to both the home agent and correspondent node to inform them of its COA. The binding acknowledgement is sent by the receiver to tell the mobile node whether the received binding update is accepted or rejected. The mobile node can specify in the binding update whether a binding acknowledgement from the receiver is required or not.

Unlike in Mobile IPv4 where the address updating messages are carried as payloads of UDP/IP packets, all three types of Mobile IPv6 address notification messages are encoded as options to be carried within an IPv6 *destinations options header.* Therefore, these messages are only examined by the ultimate destination and not by any intervening routers along the path.

Mobile IPv6 Packet Routing

When connected to their home network or link, the mobile nodes send and receive packets just as any other stationary node. When a mobile node is connected to a foreign network:

1. *For packets route from the mobile node,* the mobile node must be able to determine a router that can forward packets generated by the mobile node and then just uses standard IP routing to deliver each packet to its destination. The search for a router is easier in IPv6 than in IPv4 because all IPv6 routers are required to implement router discovery, which is not the case in IPv4. As a result, a mobile node can select any router on the foreign link from which it has received router advertisement and configures its routing table to send all packets generated by itself to that router.
2. *For packets routed to the mobile node,* if the correspondent node is aware of mobile node's current COA, it will use the mobile node's COA in the IPv6 destination address field and put the mobile node's home address in a routing header. When the mobile node receives the packets at its COA, it looks within the routing header and finds its own home address as the ultimate destination of the IPv6 packet. Thus, the mobile node consumes the packet by sending it to the higher-layer protocols.

If the correspondent node is ignorant of mobile node's COA, it puts the mobile node's home address in the IPv6 destination address field and places its own address in the IPv6 Source address field and sends the packet out. The packet will then be received by the mobile node's home agent and tunneled to the mobile node COA. Furthermore, the mobile node interprets the presence of the tunnel to mean that the correspondent node does not know the mobile node's current COA and thus will send a binding update to inform the correspondent node of its address. Once this occurs, the correspondent node will then send packets directly to the mobile node as described above (route optimization).

DISCUSSION AND CONCLUSION

Our analysis of Mobile IPv6 finds it to be better than Mobile IPv4 in several ways. It is better integrated with IPv6 than Mobile IPv4 was with IPv4, perhaps because IPv6 was designed with mobility support as one of the requirements. Route optimization allows correspondent nodes to be directly informed about the current location of the MN, so to avoid triangular routing.

However, Mobile IPv6, like all complex protocols, has its weaknesses. Researchers have found various problems. For example, Mobile IPv6 is vulnerable to certain security attacks, such as on the address notification process. An attacker that eavesdrops on the right packets could send a falsely authenticated binding update message supposedly from the MN, and thus cause traffic meant for the MN to be diverted to an arbitrary address. Preliminary works on solutions to such problems have been reported (Lim & Wong, 2005). Another example of a problem with Mobile IPv6 is its vulnerability to the "simultaneous mobility" problem. When both MN and CN are mobile, and they move simultaneously, the address notification procedure may fail. Details and proposed solutions are given in Wong and Dutta (2005).

It is expected that most of the major “holes” in Mobile IPv6 should be fixed soon, as researchers report their findings and ideas.

REFERENCES

Chirovolu, G., Agrawal, A., & Vandenhoute, M. (1999). Mobility and QoS support for IPv6-based real-time wireless Internet traffic. *IEEE Communications Magazine.*

Deering, S., & Hinden, R. (1998). *Internet protocol, version 6.* RFC 2460, IETF.

Faccin, S. M., & Le, F. (2003). A secure and efficient solution to the IPv6 address ownership problem. *Proceedings of the International Conference on Communication Technology.*

Jeong, J., Park, J., & Kim, H. (2004). Dynamic tunnel management protocol for IPv4 traversal of IPv6 mobile network. *IEEE Personal Communications Magazine.*

Johnson, D., Perkins, C.E., & Arkko, J. (2004). *Mobility support in IPv6.* RFC 3775, IETF.

Lim, E., & Wong, K. D. (2005). Binding update alternatives for Mobile IP version 6. *Proceedings of the 4th International Conference on Information Technology in Asia (CITA 2005)*, Kuching, Malaysia.

Qi, S (2001). *On providing flow transparent mobility support for IPv6-based wireless real-time services.* MEng thesis, Department of Electrical and Computer Engineering, National University of Singapore.

Samad, M., & Ishak, R. (2004). Deployment of wireless Mobile IPv6 in Malaysia. *Proceedings of the RF and Microwave Conference.*

Silva, J., Camilo, T., Costa, A., Matos, C., & Boavida, F. (2004). Exploring IPv6 mobility in IPv6 environments—issues and lessons learnt. *Proceedings of the IEEE International Conference on System, Man and Cybernatics.*

Soliman, H. (2004). *MobileIPv6—mobility on wireless Internet.* Boston: Addison-Wesley.

Wong, K. D., & Dutta, A. (2005). Simultaneous mobility in MIPv6. *Proceedings of the IEEE Electro/Information Technology Conference (EIT)*, Lincoln, NE.

KEY TERMS

Binding Acknowledgement: Same as *binding update,* but in reverse direction; acknowledgement sent by a home agent to mobile nodes, as a response to binding update.

Binding Update: Notification of a mobile node’s care-of-address to its home agent, sent by mobile nodes to a home agent.

Binding: The association between a mobile node’s home address and care-of-address.

Duplicate Address Detection (DAD): The way of checking that no node on the link is already using that address.

Neighbor Unreachability Detection (NUD): The way of checking that another host is still available and reachable on the link.

Route Optimization: Direct communication between correspondent node to any mobile node, without needing to pass through the mobile node’s home network and be forwarded by its home agent; thus eliminates the problem of triangle routing and improves routing efficiency.

Triangle Routing: Indirect communication between correspondent node and mobile node. the packet from the correspondent node will travel to its home agent first, before going to the mobile node, which is not efficient.

Enabling Multimedia Applications in Memory-Limited Mobile Devices

Raul Fernandes Herbster
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

Marcos Morais
Federal University of Campina Grande, Brazil

INTRODUCTION

Embedded systems have several constraints which make the development of applications for such platforms a difficult task: memory, cost, power consumption, user interface, and much more. These characteristics restrict the variety of applications that can be developed for embedded systems. For example, storing and playing large videos with good resolution in a limited memory and processing power mobile device is not viable.

Usually, a client-server application is developed to share tasks: clients show results while servers process data. In such a context, another hard task for limited memory/processing devices could be delegated to the server: storage of large data. If the client needs data, it can be sent piece by piece from the server to the client.

In this article we propose a layered architecture that makes possible the visualization of large videos, and even other multimedia documents, in memory/processing limited devices. Storage of videos is performed at the server side, and the client plays the video without worrying about storage space in the device. Data available in the server is divided into small pieces of readable data for mobile devices, generally JPEG files. For example, when the client requests videos from the server, the videos are sent as JPEG files and shown at an ideal rate for users. The video frames are sent through a wireless connection.

The remainder of this article is organized as follows. We begin by describing background concepts on embedded systems and client-server applications, and then present our solution to enable multimedia applications in memory-limited mobile devices. We next discuss some future trends in mobile multimedia systems, and finally, present concluding remarks.

BACKGROUND

Embedded Systems

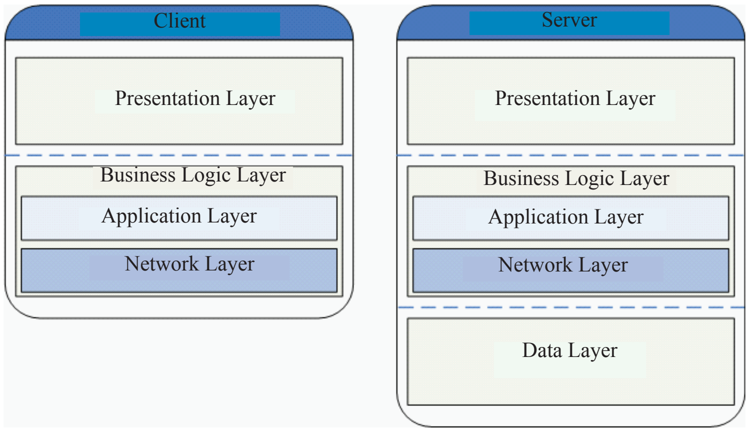
An embedded system is not intended to be a general-purpose computer. It is a device designed to perform specific tasks, including a programmable computer. A considerable number of products use embedded systems in their design: automobiles, personal digital assistants, and even household appliances (Wayne, 2005). These limited systems have some constraints that must be carefully analyzed while designing the applications for them: size, time constraints, power consumption, memory usage and disposal, and much more (Yaghmour, 2003).

These constraints restrict the variety of software for embedded systems. The development of applications which demand a large amount of memory, for example, is not viable for embedded systems, because the memory of such devices is limited. Extra memory can also be provided, but the total cost of application is very high. Another example is multimedia applications, such as video players: storing and playing large videos with good resolution in a limited memory and processing power mobile device is a very hard task.

There are specific platforms that were developed to perform multimedia tasks: embedded video decoders and embedded digital cameras, for example. However, other considerable parts of embedded systems, like personal digital assistants (PDAs) and cell phones, are not designed to play videos with good quality, store large amount of data, and encode/decode videos. Thus, it is important to design solutions enabling multimedia environments in this variety of memory/processing-limited devices.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Client/server architecture



Layered and Client-Server Architectures

Layered architectures share services through a hierarchical organization: each layer provides specific services to the layers above it and also acts as a client to the layer below (Shaw & Garlan, 1996). This characteristic increases the level of abstraction, allowing the partition of complex problems into a set of tasks easier to perform. Layered architectures also decouple modules of the software, so reuse is also more easily supported. As communication of layers is made through contracts specified as interfaces, the implementation of each module can be modified interchangeably (Bass & Kazman, 1998).

Most of the applications have three major layers with different functionalities: presentation, which handles inputs from devices and outputs to screen display; application or business logic, which has the main functionalities of the application; and data, which provides services for storing the data of the application (Fastie, 1999).

The client-server architecture has two elements that establish communication with each other: the front-end or client portion, which makes a service request to another program, called server; and the back-end or server portion, which provides service to the request. The client-server architecture allows an efficient way to interconnect programs that are distributed at different places (Jorwekar, 2005). However, the client-server architecture is more than just a separation of a user from a server computer (Fastie, 1999). Each portion has also its own modules: presentation, application, and data.

ENABLING MULTIMEDIA APPLICATIONS

Multimedia applications demand a considerable amount of resources from the environment in order to guarantee quality of service, which can be defined in terms of security, availability, or efficiency (Banâtre, 2001). Embedded systems have several constraints, like limited memory (Yaghmour, 2003), which make it very difficult to implement multimedia applications in an embedded platform.

Today, the growing interest for mobile devices and multimedia products requires the development of multimedia applications for embedded systems (Banâtre, 2001). There are approaches (Grun, Balasa, & Dutt, 1998; Leeman et al., 2005) that try to enhance embedded systems memory and other system aspects, such as processing, to provide better results in multimedia applications. However, most of the solutions available focus on hardware architecture, and a large number of programmers are not used to programming at the hardware level.

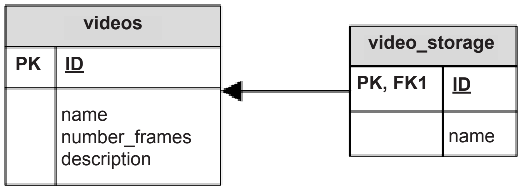
A solution based on client-server architecture is a good proposal for limited-memory/processing mobile devices because harder tasks can be performed by the server side whereas the client just displays results. By designing applications based on an architecture that shares tasks, constraints like limited memory and low computing power are partially solved. In this article, we propose a layered, client/server architecture that allows playing and storing large videos on limited-memory/processing mobile devices. The data is sent through a wireless intranet.

Client/Server Architecture

In Figure 1, both client and server modules are illustrated. Each module of both elements can be changed at any time, except the application layer because the rules of application are defined on it: if business logic changes, so does the application.

The server architecture is a standard three-tier architecture:

Figure 2. Logic model of video descriptions information



- **Presentation Layer:** This layer interacts directly with the client. Its functionalities are related to display forms so that the user adds multimedia content to the server repository.
- **Business Logic Layer:** This has two sub-layers: the network layer, which manages the connection of server and client, receiving requests and sending responses; and the application layer, which requests services for the data layer, such as document storage and reports.
- **Data Layer:** This layer stores the multimedia documents as JPEG files. Each document has information about management files, including ID, number of frames (JPEG files), and specific description elements, which depends on the multimedia document type. Figure 2 illustrates the logic model of the tables that contain such information. The server stores in a table (video_storage) the titles and ID of all videos. All the information about a video is stored in another table (videos).

The client is a single two-tier application; the data layer is not defined.

- **Presentation Layer:** This consists of a video player with buttons to select the video and a screen to display the video.
- **Business Logic Layer:** This has two sub-layers: the network layer, which manages the connection with the client, sending requests to the server and receiving data from it; and the application layer, which gets frames from the network layer and also controls the tax rate of displaying the frames.

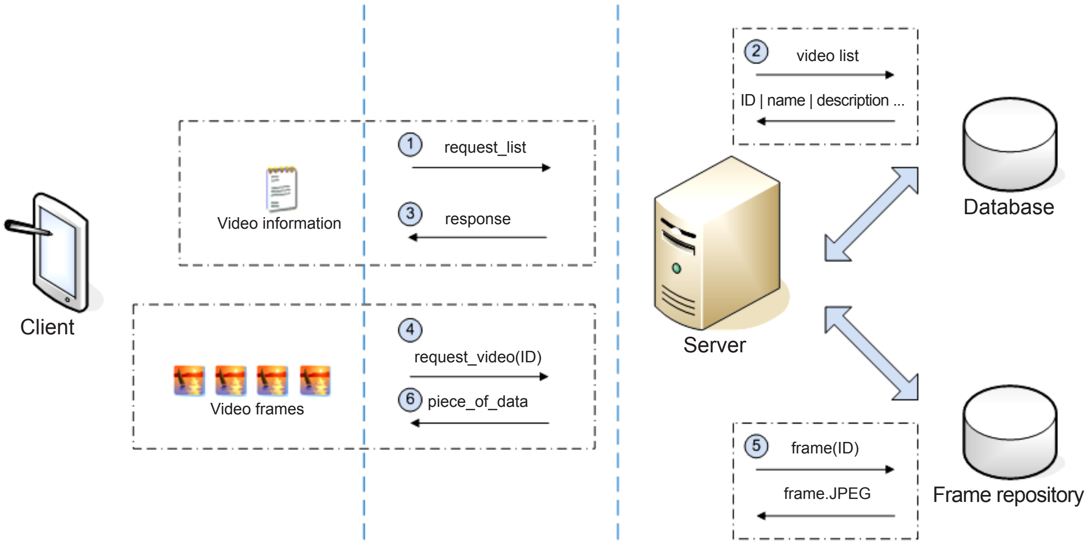
Execution Scenario

In what follows, we present an execution scenario of the mobile multimedia architecture. For this, consider that, at server side, a video was divided into small pieces of readable data (JPEG files). The quantity of pieces depends on the desired quality of the video that will be played at the client side.

The communication process between client and server is illustrated in Figure 3 and is described as follows. Whenever the server receives a connection request from a client, it sends to the client the list of available videos (steps 1 and 2). The client receives data and displays this information whenever required (step 3). Then, the client sends to the server the ID of the requested video (step 4). The server receives the request and starts the transmission of the video, piece by piece (steps 5 and 6).

At the client side, the pieces of data (JPEG files) are received in a tax rate that depends on the network (traffic, band, etc.). However, to display frames to the client in a constant tax rate and guarantee quality of service, it is necessary to maintain a buffer, which is controlled by the video player.

Figure 3. Client/server communication



In the architecture described, the data is sent through a wireless intranet. The pieces of data are JPEG files, but could also be in Motion JPEG (MJPEG) format. The tax of frames depends on the quality of the video player on the client side: generally, a high video quality rate is 30 frames per second, whereas a low video quality rate is 10 frames per second.

In a wireless network, this solution needs a large part of the network bandwidth. Therefore, there is a tradeoff between memory/processing capacity and network bandwidth. Nevertheless, considering home entertainment environments, such a tradeoff is worth the cost mainly because wireless networks in home environments have enough bandwidth to be used in such a context.

The architecture can also be used for other kinds of multimedia documents. For example, large PDF documents cannot be visualized with good quality on memory/processing-limited devices. The PDF documents can be also shared as JPEG files and sent to the client.

FUTURE TRENDS

A protocol defines communication between client and server. It is an important element to guarantee QoS. As for future work, we suggest a deeper study of protocols enabling a good service for a given situation. It is important to focus on protocols that do not demand a lot of resources from the network, such as bandwidth.

There are some protocols that are implemented over UDP, for example, trivial file transport protocol (TFTP) (RFC 783, 1981) and real-time transfer protocol (RTP) (RFC, 1996). These protocols were implemented to demand few resources of the network, and to transfer a considerable number of files through the network.

Another interesting research approach is to measure variables of the network while using an application based on the architecture described, for example, to define how many devices running such applications the network supports.

CONCLUSION

Multimedia applications demand a considerable amount of resources from systems. To develop multimedia applications for embedded systems, it is necessary to tackle constraints inherent to such platforms, such as limited memory and processing power.

In this article, we described a general architecture used for enabling multimedia applications in memory/processing systems. The architecture has two parts: a server, which receives requests and sends responses to clients; and clients, which make requests. Both parts have a layered architecture.

The solution proposed is relatively simple to implement and is easy to maintain, because the modules are decoupled and can be modified interchangeably. However, the architecture demands a considerable amount of network bandwidth, because the number of packages sent by the server to the client is large. Nevertheless, considering that the application is implemented over a wireless network in home environments, the bandwidth tradeoff is worth the cost.

REFERENCES

Banâtre, M. (2001). Ubiquitous computing and embedded operating systems design. *ERCIM News,* (47).

Bass, C., & Kazman. (1998). *Software architecture in practice.* Boston: Addison Wesley Longman.

Fastie, W. (1999). Understanding client/server computing. *PC Magazine,* 229-230.

Grun, P., Balasa, F., & Dutt, N. (1998). Memory size estimation for multimedia applications. *International Conference on Hardware Software Codesign, Proceedings of the 6th International Workshop on Hardware/Software Codesign* (pp. 145-149), Seattle, WA.

Yahgmour, K. (2003). *Building embedded Linux systems.* CA: O'Reilly.

Jorwekar, S. (2005). *Client server software architecture.*

Leeman, M., Atienza, D., Deconinck, G., De Florio, V., Mendías, J.M., Ykman-Couvreur, C., Catthoor, F., & Lauwereins, R. (2005). Methodology for refinement and optimisation of dynamic memory management for embedded systems in multimedia applications. *Journal of VLSI Signal Processing, 40*(3), 383-396.

RFC 783. (1981). *The TFTP protocol (revision 2).* Retrieved April 6, 2006, from http://www.ietf.org/rfc/rfc0783.txt?number=0783

RFC 1889. (1996). *RTP: A transport protocol for real-time applications.* Retrieved April 6, 2006, from http://www.ietf.org/rfc/rfc1889.txt?number=1889

Shaw, M., & Garlan, D. (1996). *Software architecture: Perspectives on an emerging discipline.* Englewood Cliffs, NJ: Prentice Hall.

Wolf, W. (2005). *Computer as components: Principles of embedded computing system design.* San Francisco: Morgan Kaufmann.

KEY TERMS

Client-Server Architecture: A basic concept used in computer networking, wherein servers retrieve information requested by clients, and clients desplay that information to the user.

Embedded Systems: An embedded system is a special-purpose computer system, which is completely encapsulated by the device it controls. An embedded system has specific requirements and performs pre-defined tasks, unlike a general purpose personal computer.

Embedded Software: Software designed for embedded systems.

Layered Architecture: The division of a network model into multiple discrete layers, or levels, through which messages pass as they are prepared for transmission.

Mobile Devices: Any portable device used to access a network (Internet, for example).

Multimedia Application: Applications that support the interactive use of text, audio, still images, video, and graphics.

Network Protocols: A set of rules and procedures governing communication between entities connected by the network.

Wireless Network: Networks without connecting cables, that rely on radio waves for transmission of data.

Enabling Technologies for Mobile Multimedia

Kevin Curran
University of Ulster, Northern Ireland

INTRODUCTION

Mobile communications is a continually growing sector in industry, and a wide variety of visual services such as video-on-demand have been created that are limited by low-bandwidth network infrastructures. The distinction between mobile phones and personal device assistants (PDAs) has already become blurred, with pervasive computing being the term coined to describe the tendency to integrate computing and communication into everyday life. Audio quality is highly sensitive to jitter, and video is sensitive to available bandwidth. For lip synchronization, audio and video streams need to be synchronized to within 80-100 milliseconds for skew to be imperceptible (Tannenbaum, 2005). Packets are effectively passed automatically through to the presentation device. Interpretation of the delivered information is left to human perception; because humans are far more tolerant than computers, lost packets are likely to be perceived merely as a temporary quality reduction. Nevertheless packet loss is still a significant problem for isochronous interactions. For example, since a typical packet size is generally above the threshold for audible loss (approximately 20 milliseconds), the loss of a single audio packet can be noticeable to the receiver. Resource reservation protocols are an attempt to resolve these difficulties by allocating resources prior to communication. Uncompressed multimedia data require a lot of storage capacity and very high bandwidth. Thus the use of multimedia compression is very essential. Since the source should encode the streams and the destination should decode them, multimedia compression imposes substantial loads on processing resources, such as CPU power (Yan & Mabo, 2004). New technologies for connecting devices like wireless communication and high bandwidth networks make the network connections even more heterogeneous. Additionally, the network topology is no longer static, due to the increasing mobility of users. Ubiquitous computing is a term often associated with this type of networking.

BACKGROUND

The creation of low bit rate standards such as H.263 (Harrysson, 2002) allows reasonable quality video through the existing Internet and is an important step in paving the way forward. As these new media services become available, the demand for multimedia through mobile devices will invariably increase. Corporations such as Intel do not plan to be left behind. Intel has created a new breed of mobile chip code named Banias. Intel's president and chief operating officer Paul Otellino states that "eventually every single chip that Intel produces will contain a radio transmitter that handles wireless protocols, which will allow users to move seamlessly among networks. Among our employees this initiative is affectionately referred to as 'radio free Intel'."

Figure 1. PDAs

Products such as Real Audio (www.realaudio.com) and IPCast (www.ipcast.com) for streaming media are also becoming increasingly common; however, multimedia, due to its timely nature, requires guarantees different in nature with regards to delivery of data from TCP traffic such as HTTP requests. In addition, multimedia applications increase the set of requirements in terms of throughput, end-to-end delay, delay jitter, and clock synchronization. These requirements may not all be directly met by the networks, therefore end-system protocols enrich network services to provide the quality of service (QoS) required by applications. In ubiquitous computing, software is used by roaming users interacting with the electronic world through a collection of devices ranging from handhelds such as PDAs (Figure 1) and mobile phones (Figure 2) to personal computers (Figure 3) and laptops (Figure 4).

The Java language, thanks to its portability and support for code mobility, is seen as the best candidate for such settings (Román et al., 2002; Kochnev & Terekhov, 2003). The heterogeneity added by modern smart devices is also characterized by an additional property, which is that many of these devices are typically tailored to distinct purposes. Therefore, not only memory and storage capabilities differ widely, but local device capabilities, in addition to the availability of resources changing over time (e.g., a global

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Mobiles

Figure 3. Desktops

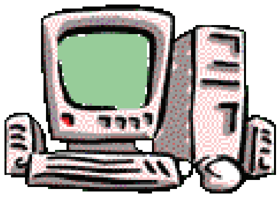


Figure 4. Laptops

positioning satellite (GPS) system cannot work indoors unless one uses specialized repeaters—see Jee, Boo, Choi, & Kim, 2003), thus a need exists for middleware to be aware of these pervasive computing properties. With regards to multimedia, applications that use group communication (e.g., videoconferencing) mechanisms must be able to scale from small groups with few members, up to groups with thousands of receivers (Tojo, Enokido, & Takizawa, 2003).

The protocols underlying the Internet were not designed for the latest cellular type networks with their low bandwidth, high error losses, and roaming users, thus many 'fixes' have arisen to solve the problem of efficient data delivery to mobile resource-constrained devices (Saber & Mirenkov, 2003). Mobility requires adaptability, meaning that systems must be location-aware and situation-aware, taking advantage of this information in order to dynamically reconfigure in a distributed fashion (Solon, McKevitt, & Curran, 2003; Matthur & Mundur, 2003). However, situations in which a user moves an end-device and uses information services can be challenging. In these situations the placement of different cooperating parts is a research challenge.

ENABLING TECHNOLOGIES FOR MOBILE MULTIMEDIA

In 1946, the first car-based telephone was set up in St. Louis in the United States. The system used a single radio transmitter on top of a tall building. A single channel was used, and therefore a button was pushed to talk and released to listen (Tanenbaum, 2005). This half-duplex system is still used by modern-day CB-radio systems used by police and taxi operators. In the 1960s the system was improved to a two-channel system, called improved mobile telephone system (IMTS). The system could not support many users, as frequencies were limited. The problem was solved by the idea of using cells to facilitate the re-use of frequencies. More users can be supported in such a cellular radio system. It was implemented for the first time in the advanced mobile phone system (AMPS). Wide-area wireless data services have been more of a promise than a reality. It can be argued that success for wireless data depends on the development of a digital communications architecture that integrates and interoperates across regional-area, wide-area, metropolitan-area, campus-area, in-building, and in-room wireless networks.

The convergence of two technological developments has made mobile computing a reality. In the last few years, the UK and other developed countries have spent large amounts of money to install and deploy wireless communication facilities. Originally aimed at telephone services (which still account for the majority of usage), the same infrastructure is increasingly used to transfer data. The second development is the continuing reduction in the size of computer hardware, leading to portable computation devices such as laptops, palmtops, or functionally enhanced cell phones. Unlike second-generation cellular networks, future cellular systems will cover an area with a variety of non-homogeneous cells that may overlap. This allows the network operators to tune the system layout to subscriber density and subscribed services. Cells of different sizes will offer widely varying bandwidths: very high bandwidths with low error rates in pico-cells, and very low bandwidths with higher error rates in macro-cells as illustrated in Table 1. Again, depending on the current location, the sets of available services might also differ.

Unlike traditional computer systems characterized by short-lived connections that are bursty in nature, Streaming Audio/Video sessions are typically long lived (the length of a presentation) and require continuous transfer of data. Streaming services will require, by today's standards, the delivery of enormous volumes of data to customer homes. For example, entertainment NTSC video compressed using

Table 1. Characteristics of various wireless networks

Type of Network	Bandwidth Latency	Latency	Mobility	Typical Video Performance	Typical Audio Performance
In-Room/Building (Radio Frequency Infrared)	>> 1 Mbps RF: 2-20 Mbps IR: 1-50 Mbps	<< 10 ms	Pedestrian	2-Way, Interactive, Full Frame Rate (Compressed)	High Quality, 16 bit samples, 22 KHz rate
Campus-Area Packet Relay	Approx. 64 kbps	Approx. 100 ms	Pedestrian	Medium Quality Slow Scan	Medium Quality Reduced Rate
Wide-Area (Cellular, PCS)	19.2 kbps	> 100 ms	Pedestrian/ Vehicular	Video Phone or Freeze Frame	Asynchronous “Voicemail”
Regional-Area (LEO/VSAT DBS)	Asymmetric Up/Dn 100 bps to 4.8 kbps 12 Mbps	>> 100 ms	Pedestrian/ Vehicular Stationary	Asynchronous Video Playback	Asynchronous “Voice Mail”

the MPEG standards requires bandwidths between 1.5 and 6 Mb/s. Many signaling schemes have been developed that can deliver data at this rate to homes over existing communications links (Forouzan, 2002). Some signaling schemes suitable for high-speed video delivery are:

- **Asymmetrical Digital Subscriber Loop:** ADSL (Bingham, 2000) takes advantage of the advances in coding to provide a customer with a downstream wideband signal, an upstream control. The cost to the end user is quite low in this scheme, as it requires little change to the existing equipment.
- **Cable TV:** CATV (Forouzan, 2002) uses a broadband coaxial cable system and can support multiple MPEG-compressed video streams. CATV has enormous bandwidth capability and can support hundreds of simultaneous connections. Furthermore, as cable is quite widely deployed, the cost of supporting video-on-demand and other services is significantly lower. However, it requires adaptation to allow bi-directional signaling in the support of interactive services.

A cellular wireless network consists of fixed based stations connecting mobile devices through a wired backbone network where each mobile device establishes contact through their local base stations. The available bandwidth on a wireless link is limited, and channels are more prone to errors. It is argued that future evolution of network services will be driven by the ability of network elements to provide enhanced multimedia services to any client anywhere (Harrysson, 2002). Future network elements must be capable of transparently accommodating and adjusting to client and content heterogeneity. There are benefits to filtering IP packets in the wireless network so that minimal application data is carried to the mobile hosts to preserve radio resources and prevent the overloading of mobile hosts with unnecessary information and ultimately wasteful processing. A proxy is an intermediary component between a source and a sink, which transforms the data in some manner. In the case of mobile hosts, a proxy is often an application that executes in the wired network to support the host. This location is frequently the base station, the machine in the wired network that provides the radio interface. As the user moves, the proxy may also move to remain on the communication path from the mobile device to the fixed network. The proxy hides the mobile from the server, which thinks that it communicates with a standard client (i.e., a PC directly connected to the wired network) (Kammann & Blachnitzky, 2002).

Wireless links are characterized by relatively low bandwidth and high transmission error rates (Chakravorty & Pratt, 2002). Furthermore, mobile devices often have computational constraints that preclude the use of standard Internet video formats on them; thus by placing a mobile transcoding proxy at the base station (BS), the incoming video stream can be transcoded to a lower bandwidth stream, perhaps to a format more suitable to the nature of the device, and control the rate of output transmission over the wireless link (Joshi, 2000).

Figure 5 illustrates a scenario where a transcoding gateway is configured to transcode MPEG streams to H.261. In the architecture, the transcoding gateway may also simply forward MPEG or H.261 packets to an alternate session (in both directions) without performing transcoding. Figure 6 illustrates locations in which intelligence about available network services may be placed. Client may utilize this network knowledge to select the most appropriate server and mechanism in order to obtain appropriate content. As an alternative, this knowledge (and the associated burden)

Figure 5. A transcoding proxy

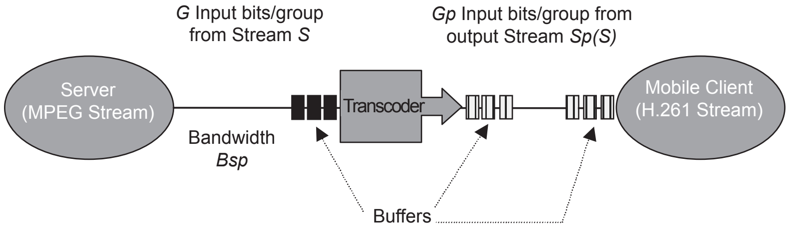
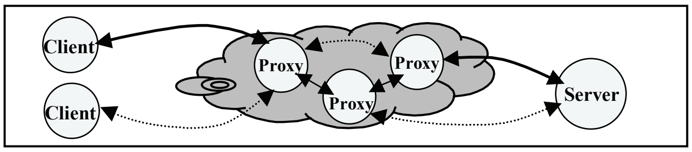


Figure 6. Variations in client-server connectivity



could be entirely or partially transferred to the individual servers or could reside inside the network.

Image transcoding is where an image is converted from one format to another (Vetro, Sun, & Wang, 2001). This may be performed by altering the Qscale (basically applying compression to reduce quality). This is sometimes known as simply resolution reduction. Another method is to scale down the dimensions of the image (spatial transcoding) (Chandra, Gehani, Schlatter Ellis, & Vahdat, 2001) so to reduce the overall byte size (e.g., scaling a 160Kb frame by 50% to 32KB). Another method known as temporal transcoding is where frames are simply dropped (this can sometimes be known as simply rate reduction), while another method may be simply to transcode the image to grayscale, which may be useful for monochrome PDAs (again this transcoding process results in reduced byte size of the image or video frame). Recently there has been increased research into intelligent intermediaries). Support for streaming media in the form of media filters has also been proposed for programmable heterogeneous networking. Canfora, Di Santo, Venturi, Zimeo, and Zito (2005) propose multiple proxy caches serving as intelligent intermediaries, improving content delivery performance by caching content. A key feature of these proxies is that they can be moved and re-configured to exploit geographic locality and content access patterns, thus reducing network server load. Proxies may also perform content translation on static multimedia in addition to distillation functions in order to support content and client heterogeneity (Yu, Katz, & Laksham, 2005). Another example is fast forward networks broadcast overlay architecture (Fast, 2005), where there are media bridges in the network which can be used in combination with RealAudio or other multimedia streams to provide an application-layer multicast overlay network. One could adopt the view at this time that "boxes" are being placed in the network to aid applications.

Padmanabhan, Wang, and Chou (2003) consider the problem of distributing "live" streaming media content to a potentially large and highly dynamic population of mobile hosts. Peer-to-peer content distribution is attractive in this setting because the bandwidth available to serve content scales with demand. A key challenge, however, is making content distribution robust to peer transience. Others (Topic, 2002) have proposed a hierarchy of retransmission servers positioned around expensive or over-utilized links. The servers operate a negative acknowledgement (NACK)-based reliable protocol between them, and receivers use a similar scheme for requesting lost packets. Their proposal significantly improves reception quality, but requires manual configuration of the retransmission servers. Streaming of stored data makes little sense unless browsing and selective playback is a requirement. For totally non-real-time scenarios, a normal transport protocol and pre-fetch can be used to achieve perfect audio quality. Media service frameworks are middleware aimed at integrating multimedia services with mobile service platforms. One such framework is PARLAY, which is the service framework of the 3rd Generation Partnership Project (3GPP, 2003). The use of the PARLAY APIs is proposed for the control of multimedia services (Parley, 2003). PARLAY is a forum established by key vendors and carriers with the goal to define object-oriented APIs that allow third-party application developers to access network resources in a generic and technology-independent way.

FUTURE TRENDS

Mobile phone technologies have evolved in several major phases denoted by "Generations" or "G" for short. Three generations of mobile phones have evolved so far, each successive generation more reliable and flexible than the previous. The first of these is referred to as the first generation or 1G. This generation was developed during the 1980s and early 1990s, and only provided an analog voice service with no data services available (Bates, 2002). The second generation or 2G of mobile technologies used circuit-based digital networks. Since 2G networks are digital, they are capable of carrying data transmissions, with an average speed of around 9.6K bps (bits per second). Because 2G networks can support the transfer of data, they are able to support Java-enabled phones. Some manufacturers are providing Java 2 Micro Edition (J2ME) (Knudsen & Li, 2005) phones for 2G networks, though the majority are designing their Java-enabled phones for the 2.5G and 3G networks, where the increased bandwidth and data transmission speed will make these applications more usable (Hoffman, 2002). These are packet based and allow for "always on" connectivity. The third generation of mobile communications (3G) (http://www.3gnewsroom.com) is digital mobile multimedia offering broadband mobile communications with voice, video, graphics, audio, and other forms of information. 3G builds upon the knowledge and experience derived from the preceding generations of mobile communication, namely 2G and 2.5G, although 3G networks use different transmission frequencies from these previous generations and therefore require a different infrastructure (Camarillo & Garcia-Martin, 2005). These networks will improve data transmission speed up to 144K bps in a high-speed moving environment, 384K bps in a low-speed moving environment, and 2Mbps in a stationary environment. 3G services see the logical convergence of two of the biggest technology trends of recent times, the Internet and mobile telephony.

Some of the services that will be enabled by the broadband bandwidth of the 3G networks include downloadable and streaming audio and video, voice-over Internet protocol (VoIP), sending and receiving high-quality color images, electronic agents that roam communications networks delivering/receiving messages or looking for information or services, and the capability to determine geographic position of a mobile device using the global positioning system (Barnes et al., 2003). 3G will also facilitate many other new services that have not previously been available over mobile networks due to the limitations in data transmission speeds. These new wireless applications will provide solutions to companies with distributed workforces, where employees need access to a wide range of information and services via their corporate intranets when they are working off-site with no access to a desktop (Camarillo & Garcia-Martin, 2005).

CONCLUSION

Flexible and adaptive frameworks are necessary in order to develop distributed multimedia applications in such heterogeneous end-systems and network environments. The processing capability differs substantially for many of these devices, with PDAs being severely resource constrained in comparison to leading desktop computers. The networks connecting these devices and machines range from GSM, Ethernet LAN, and Ethernet 802.11 to Gigabit Ethernet. Networking has been examined at a low-level micro-protocol level and again from a high-level middleware framework viewpoint. Transcoding proxies were introduced as a promising way to achieving dynamic configuration, especially because of the resulting openness, which enables the programmer to customize the structure of the system; other issues regarding mobility were also discussed.

REFERENCES

Barnes, J., Rizos, C., Wang, J., Small, D., Voigt, G., & Gambale, N. (2003, September 9-12). LocataNet: A new positioning technology for high precision indoor and outdoor positioning. *Proceedings of ION (Institute of Navigation) GPS/GNSS 2003,* Portland, OR, (pp. 1779-1789).

Bates, J. (2002, May). *Optimizing voice transmission in ATM/IP mobile networks.* New York: McGraw-Hill Telecom Engineering.

Bingham, J. (2000, January). *ADSL, VDSL, and multicarrier modulation (1st ed.)*. Wiley-Interscience.

Camarillo, G., & Garcia-Martin, M. (2005, December). *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the cellular worlds* (2nd ed.). New York: John Wiley & Sons.

Canfora, G., Di Santo, G., Venturi, G., Zimeo, E., & Zito, M. (2005). Migrating Web application sessions in mobile computing. *Proceedings of the International World Wide Web Conference* (pp. 56-62).

Chandra, S., Gehani, A., Schlatter Ellis, C., & Vahdat, A. (2001, January). Transcoding characteristics of Web images. *Proceedings of the SPIE Multimedia Computing and Networking Conference* (pp. 135-149).

Chakravorty, R., & Pratt, I. (2002, September 9-11). WWW performance over GPRS. *Proceedings of the 4th IEEE Conference on Mobile and Wireless Communications Networks (MWCN 2002),* Stockholm, Sweden, (pp. 191-195).

Feng, Y., & Zhu, J. (2001). *Wireless Java programming with J2ME* (1st ed.). Sams Publishing.

Forouzan, B. (2002). *Data communications and networking* (2nd ed.). New York: McGraw-Hill.

Harrysson, A. (2002, September 9-11). Industry challenges for mobile services. *Proceedings of the 4th IEEE Conference on Mobile and Wireless Communications Networks (MWCN 2002),* Stockholm, Sweden, (pp. 42-48).

Hoffman, J. (2002, September). *GPRS demystified* (1st ed.). New York: McGraw-Hill Professional.

Jee, G., Boo, S., Choi, J., & Kim, H. (2003, September 9-12). An indoor positioning using GPS repeater. *Proceedings of ION (Institute of Navigation) GPS/GNSS 2003,* Portland, OR, (pp. 42-48).

Kammann, J., & Blachnitzky, T. (2002, September 9-11). Split-proxy concept for application layer handover in mobile communication systems. *Proceedings of the 4th IEEE Conference on Mobile and Wireless Communications Networks (MWCN 2002)*, Stockholm, Sweden, (pp. 532-536).

Knudsen, J., & Li, S. (2005, May). *Beginning J2ME: From novice to professional.* APress.

Kochnev, D., & Terekhov, A.(2003). Surviving Java for mobiles. *IEEE Pervasive Computing, 2*(2), 90-95.

Matthur, A., & Mundur, P. (2003, September 24-26). Congestion adaptive streaming: An integrated approach. *Proceedings of DMS'2003, the 9th International Conference on Distributed Multimedia Systems,* Miami, FL (pp. 109-113).

Padmanabhan, V. N., Wang, H. J., & Chou, P. A. (2003, March). *Resilient peer-to-peer streaming.* Technical Report MSR-TR-2003-11, Microsoft Research, Redmond, WA.

Parlay Group. (2003). *PARLAY specification 2.1.* Retrieved from http://www.parlay.org

Román, M., Hess, C., Cerqueira, R., Ranganathan, A., Campbell, R., & Nahrstedt, K. (2002). A middleware infrastructure for active spaces. *IEEE Pervasive Computing, 1*(4), 74-83.

Saber, M., & Mirenkov. N. (2003, September 24-26). A multimedia programming environment for cellular automata systems. *Proceedings of DMS'2003, the 9th International Conference on Distributed Multimedia Systems,* Miami, FL (pp. 84-89).

Solon, T., McKevitt, P., & Curran, K. (2003, October 22-23). Telemorph—Bandwidth determined mobile multimodal presentation. *Proceedings of IT&T 2003,* Donegal, Ireland, (pp. 90-101).

Tanenbaum, A. (2005, May). *Computer networks* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Tojo, T., Enokido, T., & Takizawa, M. (2003, September 24-26). Notification-based QoS control protocol for group communication. *Proceedings of DMS'2003, the 9th International Conference on Distributed Multimedia Systems,* Miami, FL.

Topic, M. (2002). *Streaming media demystified.* New York: McGraw-Hill Education.

Vetro, A., Sun, H., & Wang, Y. (2001). Object-based transcoding for adaptable video content delivery. *IEEE Transactions on Circuits and Systems for Video Technology, 11*(3), 387-397.

Yan, B., & Mabo, R. (2004). QoS control for video and audio communication in conventional and active networks: Approaches and comparison. *IEEE Communications Surveys & Tutorials, 2004* (pp. 42-49). Retrieved from http://www.comsoc.org/livepubs/surveys/public/2004/jan/bai.html

Yu, F., Katz, R., & Laksham, T. (2005). Efficient multi-match packet classification and lookup with TCAM. *IEEE Micro Magazine, 25*(1), 50-59.

KEY TERMS

Bandwidth: The amount of data that can be transferred from one point to another, usually between a server and client; it is a measure of the range of frequencies a transmitted signal occupies. Bandwidth is the data speed in bits per second. In analog systems, bandwidth is measured in terms of the difference between the highest-frequency signal component and the lowest-frequency signal component.

Broadband: The telecommunication that provides multiple channels of data over a single communications medium.

Cellular Network: A cellular wireless network consists of fixed based stations connecting mobile devices through a wired backbone network where each mobile device establishes contact through their local base stations. The available bandwidth on a wireless link is limited, and channels are more prone to errors.

Encoding: Accomplishes two main objectives: (1) it reduces the size of video and audio files, by means of compression, making Internet delivery feasible; and (2) it saves files in a format that can be read and played back on the desktops of the targeted audience. Encoding may be handled by a software application, or by specialized hardware with encoding software built in.

Media: A term with many different meanings; in the context of *streaming media,* it refers to video, animation, and audio. The term "media" may also refer to something used for storage or transmission, such as tapes, diskettes, CD-ROMs, DVDs, or networks such as the Internet.

Multiple Bit Rate Video: The support of multiple encoded video streams within one media stream. By using multiple bit rate video in an encoder, you can create media-based content that has a variety of video streams at variable bandwidths. After receiving this multiple encoded stream, the server determines which bandwidth to stream based on the network bandwidth available. Multiple bit rate video is not supported on generic HTTP servers.

Streaming Video: A sequence of moving images that are transmitted in compressed form over the Internet and displayed by a viewer as they arrive; it is usually sent from pre-recorded video files, but can be distributed as part of a live broadcast feed.

Third-Generation (3G) Mobile Communications: Digital mobile multimedia offering broadband mobile communications with voice, video, graphics, audio, and other forms of information.

Enabling Technologies for Pervasive Computing

João Henrique Kleinschmidt
State University of Campinas, Brazil

Walter da Cunha Borelli
State University of Campinas, Brazil

INTRODUCTION

Bluetooth (Bluetooth SIG, 2004) and ZigBee (ZigBee Alliance, 2004) are short-range radio technologies designed for wireless personal area networks (WPANs), where the devices must have low power consumption and require little infrastructure to operate, or none at all. These devices will enable many applications of mobile and pervasive computing. Bluetooth is the IEEE 802.15.1 (2002) standard and focuses on cable replacement for consumer devices and voice applications for medium data rate networks. ZigBee is the IEEE 802.15.4 (2003) standard for low data rate networks for sensors and control devices. The IEEE defines only the physical (PHY) and medium access control (MAC) layers of the standards (Baker, 2005). Both standards have alliances formed by different companies that develop the specifications for the other layers, such as network, link, security, and application. Although designed for different applications, there exists some overlap among these technologies, which are both competitive and complementary. This article makes a comparison of the two standards, addressing the differences, similarities, and coexistence issues. Some research challenges are described, such as quality of service, security, energy-saving methods and protocols for network formation, routing, and scheduling.

BLUETOOTH

Bluetooth originated in 1994 when Ericsson started to develop a technology for cable replacement between mobile phones and accessories. Some years later Ericsson and other companies joined together to form the Bluetooth Special Interest Group (SIG), and in 1998 the specification 1.0 was released. The IEEE published the 802.15.1 standard in 2002, adopting the lower layers of Bluetooth. The specification Bluetooth 2.0+EDR (Enhanced Data Rate) was released in 2004 (Bluetooth SIG, 2004).

Bluetooth is a low-cost wireless radio technology designed to eliminate wires and cables between mobile and fixed devices over short distances, allowing the formation of ad hoc networks. The core protocols of Bluetooth are the radio, baseband, link manager protocol (LMP), logical link control and adaptation protocol (L2CAP), and service discovery protocol (SDP). The radio specifies details of the air interface, including frequency, modulation scheme, and transmit power. The baseband is responsible for connection establishment, addressing, packet format, timing, and power control. The LMP is used for link setup between devices and link management, while the L2CAP adapts upper-layer protocols to the baseband layer. The SDP is concerned with device information and services offered by Bluetooth devices.

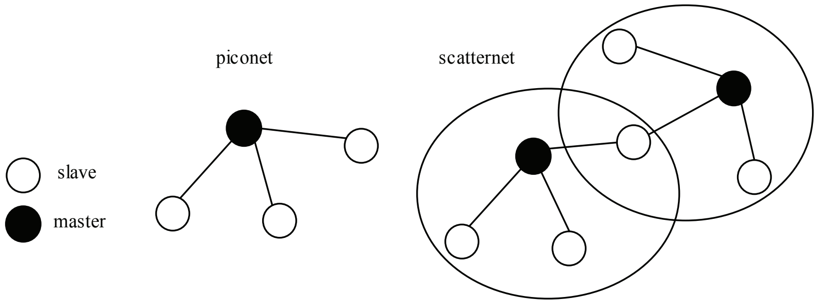
Bluetooth operates on the 2.4 GHz ISM (Industrial, Scientific, and Medical) band employing a frequency-hopping spread spectrum (FHSS) technique. There are 79 hopping frequencies, each having a bandwidth of 1MHz. The transmission rate is up to 1 Mbps in version 1.2 (Bluetooth SIG, 2003) using GFSK (Gaussian frequency shift keying) modulation. In version 2.0+EDR new modes of 2 Mbps and 3 Mbps were introduced. These modes use GFSK modulation for the header and access code of the packets, but employ different modulation for data. The $\pi/4$ differential quadrature phase-shift keying (DQPSK) modulation and 8 differential phase-shift keying (DPSK) modulation are employed in 2 Mbps and 3 Mbps mode, respectively.

The communication channel can support both data (asynchronous) and voice (synchronous) communications. The synchronous voice channels are provided using circuit switching with a slot reservation at fixed intervals. The asynchronous data channels are provided using packet switching utilizing a polling access scheme. The channel is divided in time slots of 625 μs. A time-division duplex (TDD) scheme is used for full-duplex operation.

Each Bluetooth data packet has three fields: the access code (72 bits), header (54 bits), and payload. The access code is used for synchronization and the header has information such as packet type, flow control, and acknowledgement. Three error correction schemes are defined for Bluetooth. A 1/3 rate FEC (forward error correction) is used for packet header; for data, 2/3 rate FEC and ARQ (automatic retransmission request). The ARQ scheme asks for a retransmission

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Piconet and scatternet



of the packet any time the CRC (cyclic redundancy check) code detects errors. The 2/3 rate FEC is a (15,10) Hamming code used in some packets. The ARQ scheme is not used for synchronous packets such as voice.

The devices can communicate with each other forming a network, called piconet, with up to eight nodes. Within a piconet, one device is assigned as a master node and the others act as slave nodes. In the case of multiple slaves, the channel (and bandwidth) is shared among all the devices in the piconet. Devices in different piconets can communicate using a structure called scatternet, as shown in Figure 1. A slave in one piconet can participate in another piconet as either a master or slave. In a scatternet, two or more piconets are not synchronized in either time or frequency. Each of the them operates in its own frequency-hopping channel while the bridge nodes in multiple piconets participate at the appropriate time via TDD. The range of Bluetooth devices depends on the class power, ranging from 10 to 100 meters.

ZIGBEE

ZigBee has its origins in 1998, when Motorola started to develop a wireless technology for low-power mesh networking (Baker, 2005). The IEEE 802.15.4 standard was ratified in May 2003 based on Motorola's proposal. Other companies joined together and formed the ZigBee Alliance in 2002. The ZigBee specification was ratified in December 2004, covering the network, security, and application layers (Baker, 2005).

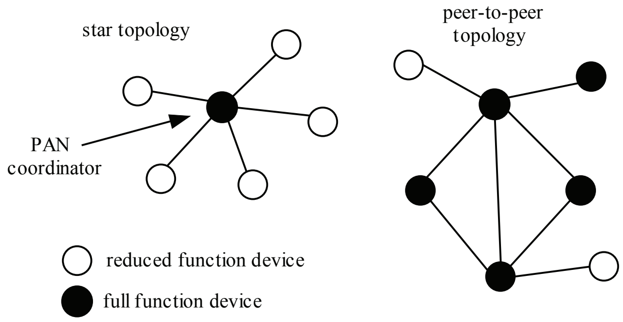
ZigBee has been designed for low power consumption, low cost, and low data rates for monitoring, control, and sensor applications (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002). The lifetime of the networks are expected to be of many months to years with non-rechargeable batteries. The devices operate in unlicensed bands: 2.4 GHz (global), 902-928 MHz (Americas), and 868 MHz (Europe). At 2.4 GHz (16 channels), the raw data rates can achieve up to 250 Kbps, with offset-quadrature phase-shift keying (OQPSK) modulation and direct sequence spread spectrum (DSSS). The 868 MHz (1 channel) and 915 MHz (10 channels) bands also use DSSS, but with binary-phase-shift keying (BPSK) modulation, achieving data rates up to 20 Kbps and 40 Kbps, respectively. The expected range is from 10-100m, depending on environment characteristics.

Each packet, called PHY protocol data unit (PPDU), contains a preamble sequence, a start of frame delimiter, the frame length, and a payload field, the PHY service data unit (PSDU). The 32-bit preamble is designed for acquisition of symbol and chip timing. The payload length can vary from 2 to 127 bytes. A frame check sequence improves the reliability of a packet in difficult conditions. There are four basic frame types: data, acknowledgement (ACK), MAC command, and beacon. The ACK frame confirms to the transmitter that the packet was received without error. The MAC command frame can be used for remote control and nodes configuration.

In 802.15.4 two channel-access mechanisms are implemented, for non-beacon and beacon network. A non-beacon network uses carrier-sense medium access with collision avoidance (CSMA-CA) with positive acknowledgements for successfully received packets. For a beacon-enabled network, a structure called superframe controls the channel access to guarantee dedicated bandwidth and low latency. The network coordinator is responsible for set up of the superframe to transmit beacons at predetermined intervals and to provide 16 equal-width time slots between beacons for contention-free channel access in each time slot (IEEE Std. 802.15.4, 2003; Gutierrez, Callaway, & Barret, 2003).

A ZigBee network can support up to 65.535 nodes, which can be a network coordinator, a full function device (FFD), or a reduced function device (RFD). The network coordinator has general network information and requires the most memory and computing capabilities of the three types. An FFD supports all 802.15.4 functions, and an RFD has limited functionalities to reduce cost and complexity. Two topologies are supported by the standard: star and peer-to-peer, as shown in Figure 2. In the star topology, the communication is performed between network devices and a single central controller, called the PAN coordinator, responsible

Figure 2. ZigBee topologies



for managing all the star functionality. In the peer-to-peer topology, every network device can communicate with any other within its range. This topology also contains a PAN coordinator, which acts as the root of the network. Peer-to-peer topology allows more complex network formations to be implemented, such as the cluster-tree. The cluster-tree network is a special case of a peer-to-peer network in which most devices are FFDs.

COMPARING ZIGBEE AND BLUETOOTH

Bluetooth and ZigBee have been designed for different applications, and this section makes a comparison between some features of both technologies, such as data rate, power consumption, network latency, complexity, topology, and scalability (Baker, 2005).

In applications where higher data rates are required, Bluetooth always has advantages, especially the 2.0+EDR version (Bluetooth SIG, 2004). While ZigBee mainly supports applications as periodic or intermittent data, achieving rates up to 250 Kbps, Bluetooth can support different traffic types, including not only periodical data, but also multimedia and voice traffic.

ZigBee devices are able to sleep frequently for extended periods to conserve power. This feature works well for energy savings, but increases the network latency because the node will have to awake in order to transmit or receive data. In Bluetooth, the devices do not sleep very often because the nodes are frequently waiting for new nodes or to join other networks. Consequently, data transmission and networks access is fast. Bluetooth devices in sleep mode have to synchronize with the network for communication, while in ZigBee this is not necessary.

The Bluetooth protocol stack is relatively complex when compared to ZigBee. The protocol stack size for ZigBee is about 28 Kbytes and for Bluetooth approximately 100 Kbytes (Geer, 2005). Bluetooth is also more complex if we consider the number of devices. A piconet can have only eight nodes, and a scatternet structure has to be formed to accommodate more nodes (Persson, Manivannan, & Singhal, 2005; Whitaker, Hodge, & Chlamtac, 2005). The Bluetooth SIG does not specify the protocols for scatternet formation. This task is easier in ZigBee networks, since no additional protocols have to be used. In terms of scalability the ZigBee also has some advantages, because network growth is easier to be implemented with flexible topologies. A Bluetooth network growth requires a flexible scatternet formation and routing protocol.

The applications have to consider these characteristics of both protocols when deciding which is the most advantageous for that specific implementation. Bluetooth will fit better in short-range cable replacement, extending LANs to Bluetooth devices and in industries for communication between fixed equipment and mobile devices or machine-to-machine communication (Baker, 2005). ZigBee is most likely to be applied in wireless sensor networks and industries wireless networks, or any other application where battery replacement is difficult and the networks have to live for months to years without human intervention. Many networks may also implement both protocols in complementary roles using the more suitable characteristic of each for that application. Table 1 shows some features of both technologies.

RESEARCH CHALLENGES

In the Bluetooth specification there is no information on how a scatternet topology should be formed, maintained, or operated (Persson et al., 2005; Whitaker et al., 2005). Two scatternet topologies that are created from separate approaches can have different characteristics. The complexity of these tasks significantly increases when moving from single piconets to multiple connected piconets.

Some research challenges in Bluetooth scatternets are formation, device status, routing, and intra and inter-piconet scheduling schemes. Each device needs to determine its role with respect to (possibly) multiple piconets, whether master and/or slave. Whitaker et al. (2005) state:

> *There is a large degree of freedom in the number of feasible alternative scatternets, which defines a significant combinatorial optimization problem. This is made more difficult by the decentralized nature of the problem, characterized by a lack of a centralized entity with global knowledge.*

The task of packet routing in a scatternet also is not so easy because the packet may have to be transmitted in multiple piconets until it reaches its destination. In a Bluetooth piconet, the master controls the channel access. A slave can send a packet only if it receives a polling packet from the master. Some slaves may participate in multiple piconets,

Table 1. Comparison between Bluetooth and ZigBee

Characteristic	**Bluetooth**	**ZigBee**
Data rate	1 Mbps (version 1.2) 3 Mbps (version 2.0)	20-250 Kbps
Expected battery duration	Days	Years
Operating frequency	2.4 GHz ISM	868 MHz, 902-928 MHz, 2.4 GHz ISM
Security	64 bit, 128 bit	128 bit AES
Network topology	Piconet and scatternet	Star, peer-to-peer, cluster tree
Protocol stack size	~100 KB	~28 KB
Transmission range	10-100 meters (depending on power class)	10-100 meters (depending on the environment)
Network latency (typical) New device enumeration Changing from sleep to active Active device channel access	 12 s 3 s 2 ms	 30 ms 15 ms 15 ms
Applications	Cable replacement, wireless USB, handset, headset	Remote control, sensors, battery-operated products

so they become more important than others and the scheduling scheme may give priority for these slaves. The nodes involved in many piconets can only be active one at a time, and the scheduling strategy has to consider this characteristic. As stated in Whitaker et al. (2005), many factors influence the design of scatternet protocols, such as distribution of devices, scalability, device differentiation, environmental dynamism, integration between coordination issues, and level of centralization. The design of efficient protocols could make Bluetooth fit for a wider range of applications.

Although in ZigBee the formation of networks with many nodes is not a great problem, the management of a network with thousands of nodes has not been addressed and may be very difficult (Geer, 2005; Zheng & Lee, 2004). Since ZigBee specification was released after Bluetooth, many important issues have not been addressed, and some distributed protocols will have to be designed for these networks. Both standards have security features, including algorithms for authentication, key exchange, and encryption, but its efficiency still has to be analyzed in networks with many nodes.

Other important issue concerning ZigBee and Bluetooth is the coexistence of both devices, as they use the same 2.4 GHz band, and channel allocation conflicts are inevitable between these WPAN technologies (Chen, Sun, & Gerla, 2006; Howitt & Gutierrez, 2003). This band is also used by wireless LANs based on IEEE 802.11 standard cordless phones and microwave ovens. Interference between near devices may be very common, so coexistence strategies have to be implemented. It is important to study the characteristics of each channel allocation scheme and how each channel allocation scheme interacts with the others. The discussion on the coexistence issue between IEEE 802.11 and the IEEE 802.15-based WPAN technologies has been included in the IEEE 802.15.2 standard.

CONCLUSION

Bluetooth and ZigBee are wireless technologies that may enable many applications of ubiquitous and pervasive computing envisioned by Weiser (1991). Millions of devices are expected to be equipped with one or both technologies in the next few years. This work addressed some of the main features and made some comparisons between them. Some research challenges were described. These issues must be properly studied for the widespread use of ZigBee and Bluetooth technologies.

REFERENCES

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y. & Cayirci, E. (2002). A survey on sensor networks. *IEEE Communications Magazine, 40*(8), 102-114.

Baker, N. (2005). ZigBee and Bluetooth: Strengths and weakness for industrial applications. *IEEE Computing and Control Engineering Journal, 16*(2), 20-25.

Bluetooth SIG. (2004). *Specification of the Bluetooth system. Core, version 2.0 + EDR.* Retrieved from http://www.bluetooth.com

Bluetooth SIG. (2003). *Specification of the Bluetooth system. Core, version 1.2.* Retrieved from http://www.bluetooth.com

Chen, L., Sun, T., & Gerla, M. (2006). Modeling channel conflict probabilities between IEEE 802.15 based wireless personal area networks. *Proceedings of the IEEE International Conference on Communications,* Istanbul, Turkey.

Geer, D. (2005). Users make a beeline for ZigBee sensor technology. *IEEE Computer, 38*(12), 16-19.

Gutierrez, J., Callaway, E., & Barret, R. (2003). *IEEE 802.15.4 low-rate wireless personal area networks: Enabling wireless sensor networks.* Institute of Electrical & Electronics Engineer (IEEE).

Howitt, I., & Gutierrez, J. (2003). IEEE 802.15 low rate-wireless personal area network coexistence issues. *Proceedings of the IEEE Wireless Communication and Networking Conference* (pp. 1481-1486), New Orleans, LA.

IEEE Std. 802.15.1. (2002). *IEEE standard for wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks.*

IEEE Std. 802.15.4. (2003). *IEEE standard for wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs).*

Persson, K. E., Manivannan, D., & Singhal, M. (2005). Bluetooth scatternet formation: Criteria, models and classification. *Elsevier Ad Hoc Networks, 3*(6), 777-794.

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 94-104.

Whitaker, R. M., Hodge, L. E., & Chlamtac, I. (2005). Bluetooth scatternet formation: A survey. *Elsevier Ad Hoc Networks, 3*(4), 403-450.

Zheng, J., & Lee, M. J. (2004). Will IEEE 802.15.4 make ubiquitous networking a reality? A discussion on a potential low power, low bit rate standard. *IEEE Communications Magazine, 42,* 140-146.

ZigBee Alliance. (2004). *ZigBee specification version 1.0.* Retrieved from http://www.zigbee.com

KEY TERMS

Carrier-Sense Medium Access with Collision Avoidance (CSMA-CA): A network contention protocol that listens to a network in order to avoid collisions.

Direct Sequence Spread Spectrum (DSSS): A technique that spreads the data into a large coded stream that takes the full bandwidth of the channel.

Frequency Hopping Spread Spectrum (FHSS): A method of transmitting signals by rapidly switching a carrier among many frequency channels using a pseudorandom sequence known to both transmitter and receiver.

Medium Access Control (MAC): A network layer that determines who is allowed to access the physical media at any one time.

Modulation: The process in which information signals are impressed on an radio frequency carrier wave by varying the amplitude, frequency, or phase.

Pervasive Computing: An environment where devices are always available and communicate with each other over wireless networks without any interaction required by the user.

Scatternet: A group of independent and non-synchronized piconets that share at least one common Bluetooth device.

Sensor Network: A network of spatially distributed devices using sensors to monitor conditions at different locations, such as temperature, sound, pressure, and so forth.

Wireless Personal Area Network (WPAN): A logical grouping of wireless devices that is typically limited to a small cell radius.

Extreme Programming for Mobile Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The liberty, expediency, and flexibility that come with mobile access have led to proliferation of mobile applications. At the same time, these applications face constant challenges posed by new implementation languages, variations in user agents, and demands for new services from user classes of different cultural backgrounds, age groups, and capabilities.

To address that, we require a methodical approach towards the development lifecycle and maintenance of mobile applications that can adequately respond to this constantly changing environment. In other words, it needs to be *agile* (Highsmith, 2002). In this article, we propose the use of an agile methodology, Extreme Programming (XP) (Beck & Andres, 2005), for a systematic development of mobile applications.

The organization of the article is as follows. We first outline the background necessary for the discussion that follows and state our position. This is followed by a discussion of the applicability and feasibility of XP practices as they pertain to mobile applications. Then the limitations of XP towards mobile applications, particularly those that are developed in an open source setting, are highlighted, and suggestions for improvement are presented. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

In recent years, ongoing efforts towards affordability of mobile devices by public-at-large and increasing contact points (service providers) have opened new vistas in the arena of mobile applications. This has also resulted in increased expectations, including sophisticated services, from mobile users. As a consequence, mobile applications continue to become increasingly large and complex. This growth, however, needs to be carefully controlled and sustained. For that, it is important that the lessons learned from the successes and failures (Nguyen, Johnson, & Hackett, 2003) in the evolution of Web applications not be ignored. Specifically, a systematic approach for creating mobile applications is desirable. We call this *Mobile Web Engineering,* inspired by traditional software engineering and Web engineering (Ginige & Murugesan, 2001).

The focus in the literature (Hjelm, 2000), however, has primarily been on implementation languages rather than the *process.* In Salmre (2005), a systematic approach to developing mobile applications is advocated, but the discussion is within the technology-specific context of Microsoft .Net Framework and Visual Basic. It is unclear how these can scale to the changing technological environment. One of the purposes of this article is to fill this gap.

We adopt the most broadly used and well-tested agile methodology, namely XP, for the development of mobile applications. XP is a test-driven "lightweight" methodology designed for small teams which emphasizes customer satisfaction and promotes teamwork. XP was created to tackle uncertainties in development environment, and in doing so, put more emphasis on the social (people) component (engineer, customer, and end user). The XP practices are set up to mitigate project risks (dynamically changing requirements, new system due by a specific timeline, and so on) and increase the likelihood of success. The use of XP has been suggested for a "rapid application development" of Web applications (Wallace, Raggett, & Aufgang, 2002; Maurer & Martel, 2002).

It is not the purpose of this article to evaluate the merit of XP on its own or with respect to other agile methodologies; such assessments have been carried out elsewhere (Turk, France, & Rumpe, 2002; Mnkandla & Dwolatzky, 2004).

ENGINEERING MOBILE APPLICATIONS USING EXTREME PROGRAMMING

In this section, we discuss in detail how the practices put forth by XP manifest themselves in the development of mobile applications (see Table 1). The 12 XP practices are: *The Planning Game, Small Releases, Metaphor Guide, Simple Design, Testing, Refactoring, Pair Programming, Collective Ownership, Continuous Integration, 40-Hour Week, On-Site Customer,* and *Coding Standards.*

We note that some of these practices such as *Testing, Refactoring*, or *Pair Programming* are not native to XP and were discovered in other contexts previously. In this sense, by aggregating them in a coherent manner, XP bases itself on "best practices." These practices are also not necessarily mutually exclusive, and we point out the relationships among them where necessary. We also draw attention to the obstacles

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. XP practices corresponding to process workflows in a mobile application

Process Workflow	XP Practices
Planning	40-Hour Week, The Planning Game (Project Velocity)
Analysis (Domain Modeling, Requirements)	On-Site Customer, The Planning Game (User Stories)
Design	Metaphor Guide (Natural Naming), Simple Design, Refactoring
Implementation	Collective Ownership, On-Site Customer, Metaphor Guide, Coding Standards, Pair Programming, Continuous Integration
Verification and Validation	On-Site Customer, Testing (Unit Tests, Acceptance Tests)
Delivery	Small Releases

in the realization of these practices that pose challenges to the deployment of XP for mobile applications.

The Planning Game

The purpose of *The Planning Game* is to determine the scope of the project and future releases by combining business priorities and technical estimates. For that, it solicits input from the "customer" to define the business value of desired features and uses cost estimates provided by the programmers. This input comes in form of *user stories* (Alexander & Maiden, 2004). A user story is a user experience informally expressed in a few lines with a mobile application such as navigating or using a search engine. The estimation is limited to the assessment of *project velocity,* a tangible metric that determines the pace at which the team can produce deliverables. The plan is prone to modifications based on the current reality.

Small Releases

The idea behind *Small Releases* is to have a simple system (an evolutionary prototype) into production early, and then via short cycles, iteratively and/or incrementally, reach the final system. To have a concrete proof-of-concept up and running can be used to solicit feedback for future versions and can help convince customers and managers of the viability of the project. This is useful for mobile applications that are highly interactive. However, there is cost associated with prototypes and therefore their number should be kept under control.

Metaphor Guide

The use of metaphors (Boyd, 1999) is prevalent in all aspects of software development. A *Metaphor Guide* is an effort to streamline and standardize efforts for naming software objects and is available for team-wide use. Natural naming (Keller, 1990) is a technique initially used in source code contexts that encourages the use of names that consist of one or more full words of the natural language for program elements in preference to acronyms or abbreviations. Indeed, natural naming strengthens the link between the underlying conceptual entity and its given name. For example, MobileProfile is a combination of two real-world metaphors placed into a natural naming scheme. The two main concerns in naming are: (1) length due to the constrained interfaces on mobile devices, and (2) user familiarity, as user background is often non-technical. For example, it is preferable to use EnterSearchWords as an indicator inside the form interface (to save space) rather than RegularExpressionForQueryString outside the form, although the latter may be a more accurate description.

Simple Design

The motivation behind a *Simple Design* is that in XP's view, requirements are *not* complete when the design commences. This is in line with the reality of mobile applications, which have to respond to the market pressures and the competition that are beyond their control, or other unavoidable circumstances such as variations in implementation technology. Therefore, the design is minimal based on *current* (not future) requirements. It aims for simplicity, and to ensure "good" design its quality (specifically, structural complexity) is improved by frequent revisitations, that is, *Refactoring.*

Testing

There is a strong emphasis in XP on validation and verification of the software at all times. By being test driven, there is transition from one phase to another only if the tests succeed. The tests range from unit tests (using tools such as HTMLUnit, HTTPUnit, XMLUnit, XSLTUnit, and JUnit) written by programmers to acceptance tests involving customers (to satisfy customer requirements). There are variations in user agents (browsers) with respect to their support for mobile markup languages, and often a single mobile device does not have the capabilities for multiple installations. Therefore, interface testing is uniquely critical to mobile applications. As part of that activity, syntactical validation of documents being served is critical. A detailed treatment of tools, techniques, and methods for testing mobile applications as well as for test plans is given in Nguyen et al. (2003).

Refactoring

The artifacts created during analysis or design may need to evolve for reasons such as discovery of "impurities" (or code "smells") or obsolescence. The refactoring (Fowler, Beck, Brant, Opdyke, & Roberts, 1999) methods are structural transformations that help eradicate the undesirables without changing the functionality of the application. Examples of such smells include inconsistent names of classes, operations, or attributes that hinder communication, redundancy (duplication), classes with unnecessary responsibility (non-cohesivity), and so on. The goal of *Refactoring* is to improve the design of the system throughout the entire development.

Pair Programming

This is one of the practices of XP that highlights the social aspects of engineering. The idea behind *Pair Programming* is to encourage collaborative work. In some controlled experiments (Williams & Kessler, 2003), Pair Programming has been shown to produce better code at similar or lower cost than programmers working alone. Empirical studies (Katira, 2004) have shown that some level of compatibility among partners in Pair Programming is necessary for it to be effective. The notion of Pair Programming can be extended to artifacts created during early stages (namely, analysis and design phases) of the development process that focus on modeling (Kamthan, 2005). For example, the use of the Unified Modeling Language (UML) (Booch, Jacobson, & Rumbaugh, 2005) for modeling mobile applications has been suggested (Grassi, Mirandola, & Sabetta, 2004). A pair can be responsible for several other practices such as using *Refactoring* to obtain a *Simple Design, Continuous Integration,* and *Testing.* The *On-Site Customer* can be a partner in a pair, but only as a co-pilot.

Collective Ownership

According to the XP philosophy, one of reasons for inertia in modifications to software is that when change is warranted, the team has to wait for specific personnel to carry it out. Therefore, in XP, all the code belongs to all the programmers and anyone can change code anywhere in the system at any time. However, for such an arrangement to be effective, configuration management that provides trace of person, date/time, and of nature and location of the change carried out is needed.

Continuous Integration

In an incremental and iterative approach of XP, standalone units (such as a corporate logo, navigation icons, and so on) are created and then integrated. However, it is not automatic that if the individual parts work, then their sum would also work. For example, a graphical navigation bar may work well individually, but may not when included in an XHTML Basic document displayed on a personal digital assistant (PDA) due to, say, incorrect encoding or link syntax. By "continuous," XP means integrating and building the software system multiple times a day. The advantage of *Continuous Integration* is minimal propagation of errors (limited to the last addition).

40-Hour Week

The term *40-Hour Week* is to be taken figuratively rather than literally. It simply implies that, due to the emphasis on the social aspect in XP, "overwork" is not recommended. XP believes that excessive overtime leads to low productivity in the long term. For example, tired programmers are prone to more mistakes, which in turn may slow down progress of the project.

On-Site Customer

The availability of a full-time *On-Site Customer* helps in understanding the application domain, determining requirements, setting priorities, and answering questions as the programmers have them. In XP, every contributor to the project, including the customer, is an integral part of the *entire* team. This has two major implications for the team: its structure is not hierarchical, and it requires physical proximity of the participants to function.

Coding Standards

There are a variety of languages for expressing information in mobile applications. The Extensible HyperText Markup Language (XHTML) is a recast of the HyperText Markup Language (HTML) in XML. XHTML Basic, a successor of Compact HTML (cHTML) and of the Wireless Markup Language (WML) 1.0, has native support for elementary constructs for structuring information like paragraphs, lists, tables, and so on. XHTML Mobile Profile (XHTML-MP) from Openwave Systems and WML 2.0 from Open Mobile Alliance (OMA) extend the functionality of XHTML Basic by adding modules as defined by the Modularization of XHTML. SVG Tiny and SVG Basic are scalable vector graphics (SVG) profiles targeted towards cellular phones and PDAs, respectively. They support two-dimensional vector graphics that work across output resolutions, across color spaces, and across a range of available bandwidths. SMIL Basic is a Synchronized Multimedia Integration Language (SMIL) profile that meets the needs of resource-constrained devices such as mobile phones and portable disc players. It allows description of temporal behavior of a multimedia presentation, associates hyperlinks with media objects, and describes the layout of the presentation on a screen. The CSS Mobile Profile is a subset of the Cascading Style Sheets (CSSs) tailored to the needs and constraints of mobile devices, and provides the presentation semantics on the client.

Information based on these languages can be served using any general or special-purpose programming language. It is critical that instances based on these languages be communicable. For example, for *Pair Programming* and for *Collective Ownership* to be effective, there needs to be a common understanding. *Coding Standards* provide means for doing that. It is known (Schneidewind & Fenton, 1996) that, when applied judiciously, standards can contribute towards quality improvement.

CHALLENGES TO THE DEPLOYMENT OF EXTREME PROGRAMMING TO MOBILE APPLICATIONS

In this section, we highlight certain caveats of applying XP practices as-is, as well as certain aspects that are essential to mobile applications but are not covered by these practices per se.

- XP does not mandate a rigorous feasibility study, including a formal means for cost estimation, as part of *The Planning Game.* A feasibility study could for instance determine if one could take advantage of reuse. For example, the functionality of a mobile application for different classes of cellular phones should not be all that different.
- The notions of *Pair Programming, 40-Hour Week,* and *On-Site Customer* make sense for a development in a non-distributed environment only. This would present a coordination obstacle if a mobile application were being developed in different natural languages, each in different parts of the world.
- Testing for usability in general and in case of mobile applications in particular can be prohibitive for small-to-medium-size enterprises, particularly if it involves specialized rooms, dedicated infrastructure with video monitoring, and recordings for subsequent analysis. Also, testing cannot always detect all errors in systems. For example, that user supplied correct address or that internal documentation corresponds to source code are beyond the scope of testing. Furthermore, testing is only one approach to verification and defect removal. XP does not include any support for formal inspections (Wiegers, 2002), although that is somewhat ameliorated by support for *Pair Programming,* which can be viewed as "informal" inspections.

XP does not explicitly take into account the licensing conditions under which the software is developed. For example, mobile applications that are open source (Kamthan, 2006) or outsourced will not be able to comply with some of the practices mandated by XP.

Finally, we point out that agility is not a panacea (Boehm & Turner, 2004). In fact, there are several issues associated with agile methodologies in general and XP in particular (Turk et al., 2002). For example, XP is not applicable to large (greater than 15) team sizes, distributed development, or for very large projects. However, XP provides a feasible first step from an ad-hoc approach to an organized view of developing mobile applications.

FUTURE TRENDS

COCOMO II (Boehm et al., 2001) provides a rigorous approach to cost (time, effort) estimation and could assist in *The Planning Game.* But that would require some adjustments in measures and calibrations of data, as mobile applications are different from traditional applications for which the COCOMO II cost estimation model is defined.

Since XP is driven by testing, there is an urgent need for unit testing frameworks (Hamill, 2004) beyond what are currently available and tailored to mobile markup languages.

For large-scale mobile applications, a "heterogeneous" process environment approach that mixes agility with discipline (Boehm & Turner, 2004) could be useful. A natural extension of the previous discussion would be to deploy a simplified version of the Unified Process (UP) (Jacobson, Booch, & Rumbaugh, 1999), which is a process *framework* that can be tailored to produce a process model for mobile applications. Indeed, such a MobileUP would on one hand be model driven, iterative, customer centric, and would on the other hand still emphasize a top-down team hierarchy and document-based communication.

CONCLUSION

Mobile applications continue to increase in size and complexity, and to sustain and manage this growth, require a systematic approach towards their development. For that, there is a need to move away from thinking at the implementation language level and focus on abstractions created *earlier* in the process. At the same time, we wish to avoid the bureaucracy in development processes that have plagued software engineering in the past.

XP provides one such viable option for development of small-to-medium-size mobile applications. The aforementioned shortcomings inherent to XP are resolvable, and pave the way towards improvements as well as considerations for other process models tailored to mobile applications.

REFERENCES

Alexander, I., & Maiden, N. (2004). *Scenarios, stories, use cases through the systems development life-cycle.* New York: John Wiley & Sons.

Beck, K., & Andres, C. (2005). *Extreme programming explained: Embrace change* (2nd ed.). Boston: Addison-Wesley.

Boehm, B. W., Abts, C., Brown, A. W., Chulani, S., Clark, B. K., Horowitz, E., et al. (2001). *Software cost estimation with COCOMO II.* Englewood Cliffs, NJ: Prentice-Hall.

Boehm, B., & Turner, R. (2004). *Balancing agility and discipline: A guide for the perplexed.* Boston: Addison-Wesley.

Booch, G., Jacobson, I., & Rumbaugh, J. (2005). *The Unified Modeling Language reference manual* (2nd ed.). Boston: Addison-Wesley.

Boyd, N. S. (1999). Using natural language in software development. *Journal of Object-Oriented Programming, 11*(9).

Fowler, M., Beck, K., Brant, J., Opdyke, W., & Roberts, D. (1999). *Refactoring: Improving the design of existing code.* Boston: Addison-Wesley.

Grassi, V., Mirandola, R., & Sabetta, A. (2004, October 11-15). A UML profile to model mobile systems. *Proceedings of the 7th International Conference on the Unified Modeling Language (UML 2004)*, Lisbon, Portugal.

Hamill, P. (2004). *Unit test frameworks: Tools for high-quality software development.* O'Reilly Media.

Hjelm, J. (2000). *Designing wireless information services.* New York: John Wiley & Sons.

Highsmith, J. (2002). *Agile software development ecosystems.* Boston: Addison-Wesley.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The Unified Software Development process.* Boston: Addison-Wesley.

Kamthan, P. (2005, January 14-16). Pair modeling. *Proceedings of the 2005 Canadian University Software Engineering Conference (CUSEC 2005)*, Ottawa, Canada.

Kamthan, P. (2006, January 19-21). Open source software in software engineering education: No free lunch. *Proceedings of the 2006 Canadian University Software Engineering Conference* (CUSEC 2006), Montreal, Canada.

Katira, N. (2004). *Understanding the compatibility of pair programmers.* MSc Thesis, North Carolina State University, USA.

Keller, D. (1990). A guide to natural naming. *ACM SIGPLAN Notices, 25*(5), 95-102.

Maurer, F., & Martel, S. (2002). Extreme programming. Rapid development for Web-based applications. *IEEE Internet Computing, 6*(1), 86-90.

Mnkandla, E., & Dwolatzky, B. (2004). A survey of agile methodologies. *Transactions of the South African Institute of Electrical Engineers, 95*(4), 236-247.

Nguyen, H.Q., Johnson, R., & Hackett, M. (2003). *Testing applications on the Web: Test planning for mobile and Internet-based systems* (2nd ed.). New York: John Wiley & Sons.

Salmre, I. (2005). *Writing mobile code: Essential software engineering for building mobile applications.* Boston: Addison-Wesley.

Schneidewind, N. F., & Fenton, N. E. (1996). Do standards improve product quality? *IEEE Software, 13*(1), 22-24.

Turk, D., France, R., & Rumpe, B. (2002, May 26-29). Limitations of agile software processes. *Proceedings of the 3rd International Conference on eXtreme Programming and Agile Processes in Software Engineering* (XP 2002) (pp. 43-46), Sardinia, Italy.

Wallace, D., Raggett, I., & Aufgang, J. (2002). *Extreme programming for Web projects.* Boston: Addison-Wesley.

Wiegers, K. (2002). *Peer reviews in software: A practical guide.* Boston: Addison-Wesley.

Williams, L., & Kessler, R. (2003). *Pair programming illuminated.* Boston: Addison-Wesley.

KEY TERMS

Agile Development: A philosophy that embraces uncertainty, encourages team communication, values customer satisfaction, vies for early delivery, and promotes sustainable development.

Coding Standard: A documented agreement that addresses the use of a formal (such as markup or programming) language.

Mobile Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering, and management principles, and disciplined and systematic approaches to the successful development, deployment, and maintenance of high-quality mobile Web applications.

Natural Naming: A technique for using full names based on the terminology of the application domain that consist of one or more words of the natural language instead of acronyms or abbreviations for elements in a software representation.

Pair Programming: A practice that involves two people such that one person (the primary person or the pilot) works on the artifact while the other (the secondary person or the co-pilot) provides support in decision making and provides input and critical feedback on all aspects of the artifact as it evolves.

Refactoring: A structureal transformation that provides a systematic way of eradicating the undesirable(s) from an artifact while preserving its behavioral semantics.

Software Engineering: A discipline that advocates a systematic approach of developing high-quality software on a large scale while taking into account the factors of sustainability and longevity, as well as organizational constraints of time and resources.

Factors Affecting Mobile Commerce and Level of Involvement

Frederick Hong Kit Yim
Drexel University, USA

Alan ching Biu Tse
The Chinese University of Hong Kong, Hong Kong

King Yin Wong
The Chinese University of Hong Kong, Hong Kong

INTRODUCTION

Driven by the accelerating advancement in information technology (IT), the penetration of the Internet and other communications services has increased substantially. Hoffman (2000), one of the most renowned scholars in the realm of Internet research, considers the Internet as "the most important innovation since the development of the printing press." Indeed, the omnipresent nature of the Internet and the World Wide Web (WWW) has been a defining characteristic of the "new world" of electronic commerce (Dutta, Kwan, & Segev, 1998). There are a good number of academics and practitioners who predict that the Internet and the WWW will be the central focus of all commercial activities in the coming decades (e.g., Dholakia, 1998). In particular, Jarvenpaa & Todd (1996) argue that the Internet is alive with the potential to act as a commercial medium and market. Figuratively, discussing the business prospects of the Internet and the WWW is somehow analogous to discussing the Gold Rush of the 19th century (Dholakia, 1995).

Admittedly, the close down of a lot of dot.coms since 2000 has been a concern for many people. However, the statistical figures we have up to now show that the growth pattern continues to be exponential. For example, the latest Forrester Online Retail Index released in January 2002 indicates that consumers spent $5.7 billion online in December, compared to $4.9 billion in November (Forrester Research, 2002a). There is yet another sign of optimism for online shopping: The Internet Confidence Index (as released in September 2002), jointly developed by Yahoo and ACNielsen, rose 13 points over the inaugural survey released in June 2001, indicating a strengthening in consumers' attitudes and confidence in e-commerce services (Yahoo Media Relations, 2002). Hence, we believe that the setback is only temporary and is part of a normal business adjustment. The future trend is very clear to us. Everybody, be it multinationals or small firms, should be convinced of the need to be on the Web.

While researchers like Sheth and Sisodia (1999) have described the growth of the Internet as astonishing, an even more startling growth is projected in the area of wireless Internet access via mobile devices. The general consensus is that mobile commerce, a variant of Internet commerce (Lucas, 2001) that lets users "surf" their phones (Wolfinbarger & Gilly, 2001), will become part of the next evolutionary stage of e-commerce (e.g., Keen, 2001; Leung & Antypas, 2001; Tausz, 2001). Mobile commerce involves the different processes of content delivery (notification and reporting) and transactions (purchasing and data entry) on mobile devices, and its current landscape resembles the Internet in its first generation in the early 1990s (Leung & Antypas, 2001). According to a study by Strategy Analytics, the rise in demand for mobile commerce services will lead to a market value of $230 billion by 2006 (Patel, 2001). Also a cause for optimism in mobile commerce services is the estimates made by the Yankee Group that the value of goods and services purchased via mobile devices will exceed $50 billion by 2005, up from $100 million in 2000 (Yankee Group, 2001). According to Yankee, the number of wireless consumers using financial services in North America alone will reach more than 35 million in 2005, a leap from the current 500,000.

Research on consumers' online behavior has so far been centered on the World Wide Web. Very few, if any, have specifically focused on mobile access despite the fact that mobile handsets are becoming increasingly popular. This is an important area of study, as the mobile phone is quickly bypassing the PC as the means of Internet access and online shopping. According to the Computer Industry Almanac, there will be an estimated 1.46 billion Internet users by 2007, compared to the 533 million today. Currently, wireless access constitutes a significant, yet limited user share of 16.0%, but by 2007, this number would have increased dramatically to 56.8% (Computer Industry Almanac, 2002). These optimistic projections are further supported by the prediction of Forrester Research that, within 5 years, up to 2.3 million wired phone subscribers in the U.S. would make the switch to

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

wireless access, making an average of 2.2 wireless phones per household by 2007 (Forrester Research, 2002b).

Aided by staggering advances in information technology, mobile devices are now capable of offering a number of Internet-based and Internet-centric services, fueling the growth of mobile commerce. The ascendancy of mobile commerce as a marketing channel warrants researchers' and practitioners' alert even in its current rudimentary stage, not only because of the huge market potential projected, but also because mobile commerce can offer new channels through which enterprises can interact with customers (Leung & Antypas, 2001). In a bid to fill the research void in the realm of mobile commerce, and to afford some insights to firms battling over the electronic commerce arena, this research was conducted with the following two objectives in mind. The first objective is to scrutinize what constitutes the weighty factors as far as transacting through mobile devices is concerned. The second one is to find out how the importance of these factors would vary when consumers are confronted with two different transactions, each with a varying degree of involvement (Celsi & Olson, 1988). The first type of transaction is a low involvement one that involves buying movie tickets with little financial commitment, while the second one is undertaking stock transactions where the stake is high.

In the following, we would briefly summarize what the literature says about important factors that affect online shopping, which forms the basis for us to speculate on factors that may be important for consumers shopping via their mobile phones, the latter being one kind of online shopping, which should resemble to some degree other forms of shopping on the Internet as far as important factors affecting consumer behavior is concerned. Hypotheses are then formulated, which is followed by the methodology. After presenting the results, we discussed the implications and conclusions of this study.

CONCEPTUALIZATION

Regardless of the mode of access, the popularity of online shopping can be partially attributed to the effectiveness and efficiency to acquire information about vendor prices and product offerings (Alba et al., 1997; Bakos, 1997; Cook & Coupey, 1998; Klein, 1998; Peterson, Balasubramanian, & Bronnenberg, 1997; Sheth, Sisodia, & Sharma, 2000; Wolfinbarger & Gilly, 2001), and convenience in overcoming geographical and time barriers (Peterson, Balasubramanian, & Bronnenberg, 1997; Sheth & Sisodia, 1999). In sum, previous literature has found that convenience, site design and financial security are dominant in determining e-satisfaction and likelihood of using the Internet as a shopping channel (Eighmey & McCord, 1998; Szymanski & Hise, 2000; Tse & Yim, 2001).

Given that mobile commerce is also one kind of online shopping, we posit that "convenience," "site design" and "financial security" are the three crucial factors affecting consumers' propensity to transact through mobile phones:

Convenience

One of the widely held perceptions that drives consumers to go online is convenience (e.g., Donthu, 1999; Wind & Mahajan, 2002). The information superhighway has been promoted as a convenient avenue for shopping (Szymanski & Hise, 2000). Driven by the growth of mobile commerce, the convenience of online shopping is further enhanced (Lucas, 2001). Li et al. (1999) find that convenience is a robust predictor of users' online buying status. Similarly, Becker-Olsen (2000) expounds that one of the most important factors that determines whether consumers buy online is the extent to which they perceive the Internet as convenient. The convenience instilled in the electronic marketplace is manifested in time savings, effort economization and accessibility, as perceived by online consumers (Wolfinbarger & Gilly, 2001). Like shopping using a PC, consumers buying movie tickets or completing stock transactions via their mobile phones would be able to save a lot of time and effort that would otherwise be wasted in dealing with agents or ticket offices.

As buying movie tickets is a transaction of low involvement and that undertaking stock transactions is of high involvement, it can be logically reckoned that the convenience factor is different in significance depending upon the situation. Convenience may have a more significant impact on consumers' propensity to transact online in the context of a ticket transaction, as compared to a stock transaction. Consumers should experience greater satisfaction when they can buy movie tickets anytime and anywhere breaking the time- and location-bound facets of traditional "gravitational" commerce (Sheth & Sisodia, 1999). On the other hand, for stock investment, consumers' major concern is security, as the consequence of any mistake can result in a great loss (Rosenbloom, 2000). Hence, we speculate that if an online stock trading system is too convenient, online investors may actually refrain from using it. For example, if, for the sake of convenience, a user is not required to enter a second password to confirm a transaction, the user may end up feeling highly insecure and less satisfied. To test our assertions, we put forward the following hypotheses:

- **H_{1a}:** Convenience significantly affects willingness to transact online for both movie ticket and stock transactions.
- **H_{1b}:** The importance of convenience in determining whether consumers transact online is greater for a ticket transaction than a stock transaction.

- **H_{1c}:** The greater the level of convenience is in an online stock transaction, the lower the intention is in using the system.

Site Design

Site design is considered important in the realm of electronic commerce (e.g., Eighmey & McCord, 1998; Lohse & Spiller, 1999; Wolfinbarger & Gilly, 2001) and mobile commerce (Lucas, 2001). Like any diffusion of innovation, there is a learning curve for most consumers to utilize electronic commerce in a way they feel most comfortable (Li, Kuo, & Russell, 1999). The success akin to the adoption of mobile transactions hinges on, at least in part, the complexity of the utilization of this innovation (Childers et al., 2001; Rogers, 1983). It can thus be understood that meticulously crafted Web sites tend to be more successful in terms of ushering in first-time online shoppers and gaining repeat visits. Empirically, Novak et al. (2000) find that a compelling online experience, which can be engendered by deliberate contrivance of Web sites, is positively associated with expected use in the future and the amount of time consumers spend online. As far as a stock transaction is concerned, we speculate that site design plays a crucial role in facilitating consumers to carry out the transaction. Users would undoubtedly require an effective design that allows them to orchestrate their portfolio without committing any mistake. Regarding the importance of site design in the use of mobile phones for ticket transactions, we undertook ten in-depth interviews with buyers who had bought tickets with their mobile phone before, and found that since most of the ticket transactions currently completed via mobile devices do not afford consumers the latitude to choose the specific seats they like because of the small screen size, they do not expect the same level of good site design as for stock transactions. Hence, our second hypothesis is formulated as below:

- **H_{2a}:** Site design significantly affects willingness to transact online for both movie ticket and stock transactions.
- **H_{2b}:** The importance of site design in determining whether consumers transact online is greater for a stock transaction than a ticket transaction.

Financial Security

It has been argued for long that the issue of financial security is pivotal in electronic transactions (e.g., Guglielmo, 1998; Kluger, 2000; Rosenbloom, 2000; Stateman, 1997), specifically in those conducted via wireless devices (Gair, 2001; Goldman, 2001; Hurley, 2001; Laughlin, 2001; Teerikorpi, 2001). As stock transactions entail a significantly larger amount of financial risk than ticket transactions, it can be logically reasoned that financial security plays a more salient role in affecting consumers' propensity to transact online when consumers are confronted with a stock transaction, as compared to a ticket transaction. Our third hypothesis is arrived at as follows:

- **H_{3a}:** Financial security significantly affects willingness to transact online for both movie ticket and stock transactions.
- **H_{3b}:** The importance of financial security in determining whether consumers transact online is greater for a stock transaction than a ticket transaction.

METHOD

Using a survey design, 192 respondents were selected by convenience sampling and interviewed in three different locations in Hong Kong representing a good cross-section of three different strata of socio-economic groups. The Hong Kong sample is deemed appropriate, by virtue of the fact that Hong Kong is the perfect location for a mobile commerce solutions player and is alive with opportunities fueled by its more than 5.77 million mobile phone subscribers, which represents a mobile penetration rate of 88.88% (Leung, 2002). The research was conducted in shopping malls, where a high customer flow can be found. Respondents were asked to make their evaluations when confronted first with a movie ticket buying scenario, and subsequently with a stock transaction using a mobile phone. For each scenario, before respondents indicate their likelihood of using the channel for the transaction, they were given a detailed description and explanation of the features and functionalities of the phone and the Web site that provide the service. Their likelihood of using the mobile service is captured by a seven-point scale ranging from "–3" to "3," where "–3" means "highly unlikely," "0" means "neither unlikely nor likely," and "3" represents "highly likely."

Each factor described above—convenience, site design, and financial security—is measured by at least four items so that the consistency/reliability of respondents' replies can be assessed. The items for each factor are based on those suggested by Keeney (1999). The factors are presented in bold type below followed by the items used for the factor. For each of the statements below, respondents were asked to indicate the extent to which they agree or disagree with it using again a seven-point scale ranging from "–3" to "3," where "–3" means "strongly disagree," "0" means "neither disagree nor agree," and "3" represents "strongly agree."

Convenience

- The mobile service is convenient.
- The mobile service can maximize transactional speed.

Table 1. Cronbach alphas for all the subscales

Factors	Ticket Transaction	Stock Transaction
Convenience	0.6236	0.8752
Site Design	0.5325	0.7122
Financial Security	0.7552	0.8870

- The mobile service can minimize waiting time.
- The mobile service can minimize personal travel.

Site Design

- The interface is easy to use.
- The interface is designed in such a way that I can contact a service staff easily if needed.
- The interface allows me to get a variety of services easily.
- The design of the interface is of high quality.
- I enjoy using the interface.

Financial Security

- The system is secure.
- Transaction conducted through the system is accurate.
- The system allows me to keep track of my previous transactions without error.
- The system provides me with clear information of my previous transactions.
- The system protects my personal financial information and privacy.

RESULTS

Table 1 reports the reliabilities of the items for each of the categories for both types of transactions: ticket and stock. All Cronbach's alpha values are greater than 0.7, except those of Convenience and Site Design for Ticket Transactions, which are marginal. However, for exploratory studies like this one, a value of Cronbach's alpha that exceeds 0.5 would be considered acceptable (Nunnally, 1978), and so the summated score of the items under each factor would be used as predictor variables for the statistical analyses described as follows:

Two regression analyses were conducted using the summated scores of the factors as the independent variables and the likelihood of transacting online (buying tickets or consummating stock transactions) as the dependent variable. The results of the regression analyses are depicted in Table 2.

Table 2. Regression coefficients for predictors of transacting online

Panel A: Regression findings for ticket transaction

Predictor Variable	Standardized Coefficient (SE)	t-value (p-level)
Convenience	0.079 (0.046)	0.787 (0.433)
Site Design	-0.122 (0.047)	-1.096 (0.275)
Financial Security	0.097 (0.034)	0.854 (0.395)

Panel B: Regression findings for stock transaction

Predictor Variable	Standardized Coefficient (SE)	t-value (p-level)
Convenience	-0.059 (0.074)	-0.274 (0.785)
Site Design	-0.328 (0.081)	-1.373 (0.175)
Financial Security	0.557 (0.061)	2.216 (0.030)

As evinced in Table 2, none of the factors significantly affect the intention to use mobile phone for buying movie tickets online. For online stock transactions, financial security exhibits the greatest impact on the likelihood of using the system, with the other two factors playing a non-significant role. Hence, our results do not lend support to $H_{1a,}$ H_{1c} and H_{2a}, and only H_{3a} is supported in so far as high involvement products are concerned.

To test the remaining hypotheses, we need to compute the difference in effect sizes for each factor in the two buying situations. The required differences, computed using the method suggested by Cohen (1977), are shown in Table 3.

The group difference effect sizes shown in Table 3 are free of the original measurement units. They measure the difference in effects of each of the factors under consideration—namely convenience, site design and financial security—on willingness to use a mobile phone to undertake the two different types of transactions. As seen in Table 3, we find that the difference in effect sizes associated with convenience is negative, which is in the expected direction. Tallmadge (1977) provides rough guidelines of difference = 0.25 indicating small effect, and difference = 0.33 for medium effect. Using this guideline, the effect size difference for convenience, is very small, so H_{1b} cannot be said to be supported although the negative direction is consistent with H_{1b}. On the other hand, the differences for site design and financial security are medium in magnitude, thus supporting H_{2b} and H_{3b}.

Table 3. Difference in effect sizes between ticket and stock transactions

	Convenience			
	Mean[1]	Standard Deviation	Pooled Within Group Standard Deviation	Difference in Effect Sizes[2]
Ticket	6.260	3.965	4.444	-0.058
Stock	6.000	5.206		
	Site Design			
	Mean	Standard deviation	Pooled within group standard deviation	Difference in Effect Sizes
Ticket	1.349	4.255	4.687	0.355
Stock	3.014	5.383		
	Security			
	Mean	Standard deviation	Pooled within group standard deviation	Difference in Effect Sizes
Ticket	1.945	6.065	6.561	0.328
Stock	4.100	7.380		

Note: [1] *Mean level of willingness to undertake respective online transactions.*
[2] *positive effect size difference is analogous to a positive treatment effect size using stock as the experimental group and ticket as the control group.*

DISCUSSION AND IMPLICATIONS

The next profound shift in the use of IT will obviously be toward wireless and mobile commerce (Keen, 2001), an emerging discipline (Varshney, 2002). The whole world of mobile commerce is about to explode (Martin, 2002). In many ways, m-commerce is, per se, the continuation of e-commerce with the Palm handheld, wireless laptops and a new generation of Web-enabled digital phones already on the market. It is even believed that portable devices such as phones, pagers and computers with mobile modems will quickly surpass desktop PCs as the Internet access devices of choice (Lindquist, 2001). The race for dominance in mobile commerce has begun (Nohria & Leestma, 2001). As addressed by Hoffman (2000), scholarly research on the Internet cannot keep abreast with business practice, let alone the scanty, if any, research on the newly emerging mobile commerce. To the very best of our knowledge, the survey we have conducted serves as a pioneer study in the realm of mobile commerce.

Aligning with our initial surmise that the salience of convenience in determining whether consumers transact online is larger for a ticket transaction than a stock transaction, our results, though only marginally supporting the hypothesis, shed light on what is deemed weighty in providing mobile commerce transactions. Enterprises providing electronic ticketing services of recreational activities, which are of low involvement, should pay heightened attention to how the convenience of their services can be enhanced. For example, the waiting time for consummating a ticket transaction as well as the transaction time required should be minimized.

Another noteworthy issue is that the coefficient associated with convenience as a predictor variable of the likelihood of consummating a stock transaction is negative. This may mean that respondents may associate increased level of convenience with increased level of inherent risks, thus hampering their propensity to transact online when the transactions in question are of high involvement. Firms facilitating online stock transactions, or other high involvement transactions, should thus be alert to this issue—they should promulgate their commitment to reduce their clients' risks in line with providing convenience.

Financial security is of paramount concern to online consumers seeking to consummate stock transactions, lending support to our third hypothesis. Indeed, the coefficient associated with financial security for stock transaction is of both practical and statistical significance (0.557; p-level = 0.030), signifying the colossal effects exerting from financial security on the likelihood to transact online. Meanwhile, the effect size difference for security is medium. The implication for our results is pronounced: financial security should be given overwhelming priority to high involvement transactions. Practically, clients should be continually and periodically informed of their online transactions, expressed in unequivocal terms. Building trust with clients is a proper and effective way to alleviate their worry about financial security (Shneiderman, 2000). This can be accomplished by nurturing and fostering a firm's relationship with its clients (Price & Arnould, 1999).

The insignificance of site design serving as a predictor of the likelihood of online transaction is contrary to what is addressed in the extant literature pertaining to traditional electronic commerce. Given the very nature of mobile commerce, we are yet afforded with some novel insights: as screens of mobile phones are miniatures of desktop monitors (Lucas, 2001), the possible designs for a site are constrained—overly fancy site designs cannot be demonstrated in the realm of mobile commerce, rendering site design insignificant in predicting transacting online. Although our study shows that site design does not play a significant role in affecting willingness to transact online, the difference in effect sizes are medium for the two different types of transactions. The latter is an indication that online firms intending to sell high involvement products should spend more on interface design when compared with their counterparts selling low involvement items.

DIRECTIONS FOR FUTURE RESEARCH

As the costs pertaining to the two online transactions delineated previously are somehow kept constant, we have not yet examined the effects of costs on propensity to transact online. The issue should be addressed in future research, since minimizing costs is identified as an objective in online transactions (Keeney, 1999; Leavy, 1999).

The intriguing result of the negative coefficient associated with convenience in the realm of high involvement transactions also warrants further research. Not until forthcoming research renders our intuitive interpretation to empirical scrutiny will a more comprehensive understanding of mobile commerce ensue.

Furthermore, other types of products should be chosen in addition to the ones we have used in this research to improve the generalizability of our findings. Meanwhile, in addition to classifying products based on level of involvement, we may categorize products in other ways. For example, we may classify products as search, experience and credence goods (Klein, 1998), and the relative importance of the three factors studied may be different for these three categories of products.

REFERENCES

Alba, J., Lynch, J., Weitz, B., Janiszewski, C., Lutz, R., Sawyer, A., & Wood, S. (1997, July). Interactive home shopping: Consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *Journal of Marketing, 61,* 38-53.

Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronics marketplaces. *Management Science, 43*(12), 1676-1692.

Becker-Olsen, K. L. (2000). *Point, click and shop: An exploratory investigation of consumer perceptions of online shopping.* Paper presented at AMA summer conference.

Celsi, R. L., & Olson, J. C. (1988). The role of involvement in attention and comprehension processes. *Journal of Consumer Research, 15,* 210-224.

Childers, T. L., Carr, C. L., Peck, J., & Carson, S. (2001). Hedonic and utilitarian motivations for online retail shopping behavior. *Journal of Retailing, 77*(4), 511-535.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Computer Industry Almanac. (2002). *Internet users will top 1 billion in 2005. Wireless Internet users will reach 48% in 2005.* Retrived December 31, 2005, from http://www.c-i-a.com/pr032102.htm

Cook, D. L., & Coupey, E. (1998). Consumer behavior and unresolved regulatory issues in electronic marketing. *Journal of Business Research, 41,* 231-238.

Dholakia, R. R. (1995). *Connecting to the Net: Marketing actions and market responses*. Paper presented at the International Seminar on Impact of Information Technology hosted by CIET-SENAI, Rio de Janeiro, Brazil, December 6, 1995.

Dholakia, R. R. (1998). Special issue on conducting business in the new electronic environment: Prospects and problems. *Journal of Business Research, 41,* 175-177.

Donthu, N. (1999). The Internet shopper. *Journal of Advertising Research, 39*(3), 52-58.

Dutta, S., Kwan, S., & Segev, A. (1998). Business transformation in electronic commerce: A study of sectoral and regional trends. *European Management Journal, 16*(5), 540-551.

Eighmey, J., & McCord, L. (1998). Adding value in the information age: Uses and gratifications of sites on the World Wide Web. *Journal of Business Research, 41,* 187-194.

Forrester Research. (2002a). *December shopping up from last year in spite of rough economy, according to the Forrester Research Online Retail Index*. Retrieved December 31, 2005, from http://www.forrester.nl/ER/Press/Release/0,1769,678,00.html

Forrester Research. (2002b). Retrieved from http://www.forrester.com

Gair, C. (2001). The next big thing? *Black Enterprise, 31*(10), 62.

Goldman, C. (2001). Banking on Security. *Wireless Review, 18*(7), 22-24.

Guglielmo, C. (1998). Security fears still dog Web sales. *Inter@ctive Week, 5*(273), 44-47.

Hoffman, D. L. (2000). The revolution will not be televised: Introduction to the special issue on marketing science and the Internet. *Marketing Science, 19*(1), 1-3.

Hurley, H. (2001). Pocket-sized security. *Telephony, 240*(18), 42-50.

Jarvenpaa, S. L., & Todd, P. A. (1996). Consumer reactions to electronic shopping on the World Wide Web. *International Journal of Electronic Commerce, 1*(2), 59-88.

Keen, P. G. W. (2001). Go mobile—now! *Computerworld, 35*(24), 36.

Keeney, R. L. (1999). The value of Internet commerce to the customer. *Journal of the Institute for Operations Research and the Management Sciences, 45*(4), 533-542.

Klein, L. R. (1998). Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of Business Research, 41,* 195-203.

Kluger, J. (2000). Extortion on the Internet. *Time, 155*(3), 56-58.

Laughlin, K. (2001). Banking on wireless. *America's Network, 105*(1), 56-60.

Leavy, B. (1999). Organization and competitiveness—Towards a new perspective. *Journal of General Management, 24*(3), 33-52.

Leung, K., & Antypas, J. (2001). Improving returns on m-commerce investments. *Journal of Business Strategy, 22*(5), 12-13.

Leung, T. (2002, May 27). HK trails in mobile data. *Asia Computer Weekly,* 1.

Li, H., Kuo, C., & Russell, M. G. (1999). The impact of perceived channel utilities, shopping orientations, and demographics on the consumer's online buying behavior. *Journal of Computer Mediated Communication, 5*(2).

Lindquist, C. (2001). Mobile Internet access exploding. *CIO, 14*(13), 138.

Lohse, G. L., & Spiller, P. (1999). Internet retail store design: How the user interface influences traffic and sales? *Journal of Computer Mediated Communication, 5*(2).

Lucas, P. (2001). M-commerce gets personal. *Credit Card Management, 14*(1), 24-27.

Martin, N. (2002). Content a la Wmode: Serving up solutions for wireless content. *EContent, 25*(1), 48-49.

Nohria, N., & Leestma, M. (2001). A moving target: The mobile-commerce customer. *MIT Sloan Management Review, 42*(3), 104.

Novak, T. P., Hoffman, D. L., & Yung, Y. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing Science, 19*(1), 22-42.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Patel, N. (2001). *Mobile commerce market update.* Retrieved December 31, 2005, from http://www.strategyanalytics.net/default.aspx?mod=ReportAbstractViewer&a0=839

Peterson, R. A., Balasubramanian, S., & Bronnenberg, B. J. (1997). Exploring the implications of the Internet for consumer marketing. *Journal of the Academy of Marketing Science, 25*(4), 329-346.

Price, L. L., & Arnould, E. J. (1999). Commercial friendships: service provider-client relationships in context. *Journal of Marketing, 63,* 38-56.

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed.). New York: Free Press.

Rosenbloom, A. (2000). Trusting technology. *Communications of the ACM, 43*(12), 31-32.

Sheth, J. N., & Sisodia, R. S. (1999). Revisiting marketing's lawlike generalizations. *Journal of the Academy of Marketing Science, 27*(1), 71-87.

Sheth, J. N., Sisodia, R. S., & Sharma, A. (2000). The antecedents and consequences of customer-centric marketing. *Journal of the Academy of Marketing Science, 28*(1), 55-66.

Shneiderman, B. (2000). Designing trust into online experiences. *Communications of the ACM, 43*(12), 57-59.

Stateman, A. (1997). Security issues impact online buying habits. *Public Relations Tactics, 4*(10), 8-16.

Szymanski, D. M., & Hise, R. T. (2000). E-satisfaction: An initial examination. *Journal of Retailing, 76*(3), 309-322.

Tallmadge, G. K. (1977). *The Joint Dissemination Review Panel ideabook.* Washington, DC: National Institute of Education and U.S. Office of Education.

Tausz, A. (2001). Customizing your world. *CMA Management, 75*(2), 48-51.

Teerikorpi, E. (2001). How secure is the wireless Internet. *Telecommunications, 35*(5), 46-47.

Tse, A. C. B., & Yim, F. (2001). Factors affecting the choice of channels: Online vs. conventional. *Journal of International Consumer Marketing, 14*(2/3), 137-152.

Varshney, U. (2002). Multicast support in mobile commerce applications. *Computer, 35*(2), 115-117.

Wind, Y., & Mahajan, V. (2002). Convergence marketing. *Journal of Interactive Marketing, 16*(2), 64-79.

Wolfinbarger, M., & Gilly, M. C. (2001). Shopping online for freedom, control, and fun. *California Management Review, 43*(2), 34-55.

Yahoo Media Relations. (2001). *Internet confidence index.* Retrieved December 31, 2005, from http://docs.yahoo.com/docs/info/yici/06-02.html

Yankee Group. (2001). Retrieved from http://www.yankeegroup.com

KEY TERMS

Convenience: One of the determining factors for e-satisfaction and likelihood of using the Internet as a shopping channel. It is manifested in time savings, effort economization and accessibility, as perceived by online consumers.

Financial Security: One of the determining factors for e-satisfaction and likelihood of using the Internet as a shopping channel. It refers to the personal financial information protection for the consumers who make online transactions.

Involvement: A consumer's overall subjective feeling of personal relevance.

Mobile Commerce: A variant of Internet commerce that lets users "surf" their mobile devices, for example, mobile phone, PDA.

Online Shopping: Transactions made via Internet rather than at a physical location by consumers.

Site Design: One of the determining factors for e-satisfaction and likelihood of using the Internet as a shopping channel. It refers to the interface quality that a company provides for its consumers to do online transactions.

A Game-Based Methodology for Collaborative Mobile Applications

Michael Massimi
University of Toronto, Canada

Craig H. Ganoe
The Pennsylvania State University, USA

John M. Carroll
The Pennsylvania State University, USA

INTRODUCTION

Mobile computing, perhaps more so than traditional desktop computing, requires methods for allowing application designers to try ideas, create prototypes, and explore the problem space. This need can be met with rapid prototyping. Rapid prototyping is a technique that permits members of a design team to iterate through several versions of their low-level designs (Thompson & Wishbow, 1992). During each cycle of each prototype, the design team identifies critical use cases, verifies requirements are being met, and gathers both subjective and objective data regarding usability. Because "shallow" or low-fidelity prototypes can be quickly created, used, and thrown away (Sefelin, Tscheligi, & Giller, 2003), the team can explore many options and designs with less effort than it would take to create "deep" or high-fidelity versions of each prototype (Rudd, Stern, & Isensee, 1996).

Rapid prototyping techniques are especially valuable when the application is intended for a mobile user. This is for three primary reasons. First, the mobile user is likely to be simultaneously attending to a dynamic or unpredictable environment. This environment taxes the user's cognitive abilities. Users must navigate to their destinations, avoiding obstacles and responding to changing conditions. Non-technical aspects can change, like weather or available routes. Many times, the user must "make place" in order to use the system, stopping to seek out an area to use the software (Kristoffersen & Ljunberg, 1999). Technical aspects of the system, such as network availability and power levels, can also be difficult to accurately predict and may require complex adaptation algorithms (Noble et al., 1997; de Lara, Kumar, Wallach, & Zwaenepoel, 2003; Welch, 1995). Compared to a stationary environment, the number of things that can go wrong seems to skyrocket.

Second, interpersonal communication changes when a dimension of mobility is introduced. When working collaboratively on a task, users require awareness of the tasks their collaborators are performing (Ganoe et al., 2003) in order to prevent redundancy and achieve an equitable distribution of work. When users are mobile, however, awareness is no longer simply *what* other people are doing, but also *where* they are doing it. This introduces a need for additional application support for mobile collaborative systems.

Third, heterogeneity of devices results in different interaction styles. Mobile phones provide an excellent example of this problem. Each manufacturer repositions buttons based on hardware and space constraints. Even within a manufacturer's own product line, multiple key configurations occur. This is to say nothing of the variety of mobile devices available—PDAs, tablet PCs, wearable computers, and so on. Some of the large manufacturers, like Palm, provide human interface guidelines to third-party developers (Ostrem, 2003). Most do not.

In terms of evaluating systems, Abowd and Mynatt (2000) argue that our current methods are not sufficient. The traditional task-based evaluation methods no longer apply in a world where we cannot always experimentally control the environment, and where there is not a clear, single indicator of task performance. There are not established tests that can be performed to determine the effectiveness of deployed systems, mainly because there are not many of them in the world yet. Because we do not have a base of knowledge regarding how to design for mobile interaction, early affirmations of whether the application will serve a human need are critical, and Abowd and Mynatt state that we should "understand how a new system is used by its intended population before performing more quantitative studies on its impact" (p. 47).

Mobile systems need fast, inexpensive ways of prototyping and gathering usability results. This entry describes previous work in rapid prototyping for mobile systems. We then contribute a novel rapid prototyping methodology for mobile systems, which we call "Scavenger Hunt." It is anticipated that this methodology will be useful not only for those interested in rapid prototyping and design methodologies, but also for design teams with real deadlines to meet. Finally, we identify future trends in prototype evaluation of mobile systems.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

BACKGROUND

Games

Our prototype evaluation methodology is based on a game—specifically, a Scavenger Hunt. The basis for this choice stems from success with using games as a tool for design and testing for non-mobile applications.

Twidale and Marty (2005) used a "game show" format during a conference, wherein contestants found usability problems in software, cheered on by an audience. They argue that "it is worth exploring the power of rapid, lightweight methods to catch relatively uncontroversial and easily fixed usability flaws." Scavenger Hunt does this as well, although the focus of the participant is not on the actual discovery of the flaw, but on completing a higher-level task.

Spool, Snyder, Ballman, and Schroeder (1994) created a game where designers are placed onto teams and are given a time limit to create a UI. Then, test users move from design to design and must complete the same task on each one. The design with the quickest task completion time is the winner. Here, the goal is to teach designers how to create usable software by rewarding them in a game. In this study, the game is used educationally. The goal of the game is to teach the player how to create good designs, or how to use a particular evaluation method (e.g., heuristic evaluation). Instead, we use a game itself to *evaluate* the prototype. This game-based evaluation is designed to compliment other lightweight usability evaluation metrics like heuristic evaluations (Nielsen & Molich, 1994).

Pedersen and Buur (2000) created a board game to help participatory design teams conceptualize their sessions. The board, modeled after the industrial plant where the users worked, was populated with foam pieces representing artifacts and people. The design partners took turns moving the pieces to explain processes in the plant, and this opened the door to discussion about what should and should not occur during a particular process. The notion of turn-taking is especially noteworthy, as it allows design partners to offer their thoughts and obtain equal footing in the design process. We move from a board game to a "real-life" game in the SH process. In addition, we are interested in using a game as an evaluation tool rather than a design tool. Despite these differences, the past successes with games as parts of the design lifecycle are very encouraging.

Mobile Design and Usability

In experiments conducted by Virzi, Sokolov, and Karis (1996), it was found that testing with low-fidelity prototypes found almost as many usability problems as their high-fidelity counterparts. We argue, however, that paper prototypes will not be suitable for mobile interaction, and that low-fidelity computer-based versions of prototypes should be used instead.

SCAVENGER HUNT

Motivation

To gather usability metrics about mobile collaboration systems, we have developed a methodology we call "Scavenger Hunt" (SH). SH emulates the children's game where players are given a list of items that they must collect and bring to a pre-ordained location. In our methodology, the "players" are in fact target users, and each is equipped with the appropriate mobile device and prototype software under scrutiny.

By basing the rapid prototyping technique on a well-known game, the users can quickly be brought up to speed on how to complete the usability test. Further, they are motivated to "win" the game by completing all the tasks to the best of their ability. This combats the ennui that might otherwise set in when a user is simply asked to perform a series of artificial tasks. In fact, a savvy usability tester might pit two teams against one another to see who wins first and by what methods. Extreme use cases are more likely to emerge when users push the system to its boundaries to win.

Study Details

We conducted a pilot study wherein we used the SH method to evaluate a collaboration tool prototype. The specific details we have used to conduct this SH session follow and are meant to serve as an early model for future applications of this method. These details and parameters can, of course, be tailored to meet the needs of a particular design team, product, or schedule.

Software

In order to pilot the Scavenger Hunt method, we developed a Weblog prototype as the software under scrutiny. The Weblog (which we call SH Blog) allowed multiple people to add posts to it, edit each others' posts, and reorganize the ordering of the posts. We purposefully did not create a "polished" version of the software. The prototype was representative of a first pass through coding the system and was written in approximately five hours.

The prototype was written in PHP and HTML. Clients ran Microsoft Internet Explorer for Pocket PC and rendered pages from an Apache Web server running on Linux. Data was stored server-side in a MySQL relational database.

Participants

Eleven participants were recruited from a summer school program for gifted youth. They were divided into groups of three (one participant failed to arrive). Each team was self-selected and worked together on a project during the summer school program, so the participants were comfortable working with each other. Overall they reported high levels of comfort with technology, but did not use mobile computers very often.

Pre-Session Setup

Before the SH session, we distributed 24 clues throughout the building. These clues were evenly divided among the three floors of the building. All clues were printed on brightly colored paper and were hung on walls or placed on tables. We attempted to disperse the clues throughout the building evenly so that a participant would have a chance to find a clue in consistent time intervals (e.g., after about 45 seconds of walking). We ensured that all clues were in public areas so that participants would have access to them and would not feel awkward entering private offices.

The SH Blog was engineered to capture data about user interactions before the session. The time of posting and user who posted were logged. A software engineer monitored the MySQL database that stored SH blog posts and noted the progress of the team. This monitoring was essential to the evaluation of the prototype from the software engineering perspective, as it allowed us to look "under the hood" of the software during the session.

Finally, we ensured IEEE 802.11b wireless networking was available in all areas where there were clues. Some areas, such as stairwells or elevators, could not receive a signal; this is characteristic of most mobile computing environments, however.

Starting an SH Session

We gathered each team individually at the beginning of the session and had them complete a questionnaire that asked questions about their comfort level with mobile computers, their experience with working while mobile, and their preferences for group work. Each team was then given an overview of the game that explained the following:

- There are clues throughout the building. They are all in plain view and are in public spaces.
- You need to collect as many clues as you can in order to answer a riddle.
- You must use the SH Blog to share the clues and to work on solving the riddle. You may not use software on your mobile computer besides the SH Blog.
- You will have one hour to complete the task and return here with the answer to the riddle.

Participants were then given the riddle and asked to begin. They immediately began the task and started to walk around the building, entering clues into the SH Blog.

Collecting Data During an SH Session

During the study, participants were videotaped by researchers with camcorders in order to later analyze comments and note salient themes. Participants were asked to think aloud in order to capture the cognition accompanying the interaction and problem solving. At the end of the session, users completed a questionnaire about their experience with the SH Blog prototype. Finally, based on observations during the task and questionnaire responses, we conducted a brief semi-structured interview wherein we asked questions about problems, ideas for changes, and experiential preferences. By using three different research instruments, we are able to collect a wide range of data and make design suggestions based on both explicit and implicit behaviors.

At the system level, we captured information about the number of posts made by each team member, the total number of posts, the movement of posts, the deletion of posts, and so on. By charting these over time and comparing the different groups, we can note differences in usage and system support. For example, one group in our study generated a new post for each clue they discovered; another chose to accumulate all clues in a single post. These varying styles indicate, for example, the need for the SH Blog to accommodate both large numbers of posts and large, monolithic posts.

Evaluation and Results

Some of the salient results of our trial run are presented. It is important to note the different types of problems that the method identified—they run the gamut from usability issues, to systems issues, to social issues. Many of these insights may not have been found by traditional task-based user tests.

We noted that no group collected the entire set of clues. This was for a variety of reasons. A team member might believe that a different team member already collected the clue. A team might miss the clue completely. A team might have found it and subsequently deleted it, reasoning that it was redundant or useless. Without a sufficient subset of the clues collected, teams could not solve the problem.

Different teams approached the game differently, even though they were all given the same starting conditions. One team chose to divide the building into floors and then assigned a floor to each member. Another team chose to have one member act as an analyst back at the "base," while the other two members walked around and focused solely on the collection aspect. This indicates that high-level "gaming" strategies must be supported in the software.

In every trial, the members initially split up and went separate ways to find clues. Again, in every trial, the members eventually met face-to-face once they thought they had collected all of the clues. We noticed a shift in work styles from an individual, mobile worker to a stationary, group worker. This indicates that the software must accommodate individual work and group work separately, and provide a transition between the two work styles.

Of the four trials, only one team actually solved the problem. Even this team had boiled it down to an educated guess. Based on this outcome and the ones given above, we determined that the SH Blog did not allow the users to accomplish the task and needed revisions. The ability to reorder posts more easily and the ability to draw free-form tables were the primary revisions that users identified in the questionnaires and interviews. In one exceptional case, a user accidentally deleted the aggregate post that contained all of their collected clues! It was only then that we realized there was no undo function.

As these four themes suggest, users identified numerous flaws and areas for improvement in the software during interviews, questionnaires, and observations. Because the goal of this study was for feasibility purposes, we have not evaluated our method against other methods on similar tasks. We do, however, feel that the insights gained from using SH were well worth the setup costs. The major contribution of using SH is to show that rapid, low-cost usability and systems testing can be conducted early in the design process. This method may be of use to design teams in both research and industry settings.

FUTURE TRENDS

As mobile applications are developed more frequently in order to suit the needs of the third-wave information worker, we believe that prototyping, iterative design, and usability testing will become more and more important. Cell phones demonstrate that users prefer mobile devices when they are used for interpersonal contact, and methods for evaluating collaboration on-the-go are essential for this task. Techniques like SH are useful for software engineers and usability engineers alike. We believe more tools like it should be developed.

Further, the evaluation of these tools is an open research problem. How do we demonstrate that one mobile system suits its users' requirements more effectively than another? What are the outcomes to be measured? In the absence of a long history of software deployment and use, these questions remain to be answered.

CONCLUSION

As Abowd and Mynatt (2000) observed, it is extremely difficult to conduct evaluations of mobile computing systems because of the always-on, dynamic environment. For this reason, it is imperative that we have tools for early, non-trivial user testing, and we have presented a novel method for doing so. Our method is lightweight and can be applied repeatedly in the design process to ensure that requirements are met before an expensive deployment begins. Although it does not replace actual field trials, it can identify systems-level and interface-level flaws by simulating a representative task in the problem domain. Continued work in identifying the critical components of evaluating mobile systems is important, as is the need for early prototyping and validation.

REFERENCES

Abowd, G., & Mynatt, E. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction, 7*(1), 29-58.

de Lara, E., Kumar, R., Wallach, D. S., & Zwaenepoel, W. (2003). Collaboration and multimedia authoring on mobile devices. *Proceedings of the 1st International Conference on Mobile Systems, Applications, and Services (MobiSyS)* (pp. 287-301).

Ganoe, C., Somervell, J., Neale, D., Isenhour, P., Carroll, J., Rosson, M. B., et al. (2003). Classroom BRIDGE: Using collaborative public and desktop timelines to support activity awareness. *Proceedings of the 16th ACM Symposium on User Interface Software and Technology (UIST)* (pp. 21-30).

Kristoffersen, S., & Ljunberg, F. (1999). Making place to make it work: Empirical exploration of HCI for mobile CSCW. *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work 1999* (pp. 276-285).

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems 1990* (pp. 373-380).

Noble, B., Satyanarayanan, M., Narayanan, D., Tilton, J. E., Flinn, J., & Walker, K. (1997). Agile application-aware adaptation for mobility. *Proceedings of the 16th ACM Symposium on Operating Systems Principles* (pp. 276-287).

Ostrem, J. (2003). *Palm OS user interface guidelines.* Retrieved January 6, 2006, from http://www.palmos.com/dev/support/docs/ui/UIGuidelinesTOC.html

Pedersen, J., & Buur, J. (2000). Games and movies—Towards innovative co-design with users. *Proceedings of the CoDesigning Conference,* Coventry, UK.

Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *ACM Interactions, 3*(1), 76-85.

Sefelin, R., Tscheligi, M., & Giller, V. (2003). Paper prototyping—What is it good for? A comparison of paper- and computer-based low-fidelity prototyping. *CHI 2003 Extended Abstracts,* 778-779.

Spool, J. M., Snyder, C., Ballman, D., & Schroeder, W. (1994). Using a game to teach a design process. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 117-118).

Thompson, M., & Wishbow, N. (1992). Prototyping: Tools and techniques: Improving software and documentation quality through rapid prototyping. *Proceedings of the 10th Annual International Conference on Systems Documentation* (pp. 191-199).

Twidale, M., & Marty, P. (2005). Come on down! A game show approach to illustration usability evaluation methods. *IEEE Interactions, 12*(6), 24-27.

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 236-243).

Welch, G. (1995). A survey of power management techniques in mobile computing operating systems. *ACM SIGOPS Operating Systems Review, 29*(4), 47-56.

KEY TERMS

Evaluation Methodology: A procedure for determining the quality of a system in relation to how it satisfies user needs. Scavenger Hunt is an example of an evaluation methodology for rapidly prototyped mobile collaboration systems.

Extreme Use Case: Unlike a "critical use case" where a task is identified that is essential to the operation of the system, an "extreme use case" is the situation that arises when users interact with the software under stressful (i.e., timed) conditions and push the software to its limits, both in terms of system-level support and usability.

Mobile Collaboration: The situation that arises when two or more people must work together on a problem, while one or more of them is in the process of changing location or is in the field.

Participatory Design: The process of designing *with* users instead of designing *for* users, by actually including end users on the design team and mutually learning from one another.

Rapid Prototyping: The process of creating, evaluating, and refining low-cost, easily fabricated prototypes in order to quickly identify and fix flaws.

Scavenger Hunt: A lightweight method for evaluating prototypes early in the design of a mobile or ubiquitous computing system, wherein participants play a game while the design team identifies systems- and interface-level flaws.

SH Blog: A collaborative mobile Weblog that is shared among a group of people working on the same task. The people who are involved are also posting from mobile devices.

Gender Difference in the Motivations of Mobile Internet Usage

Shintaro Okazaki
Autonomous University of Madrid, Spain

INTRODUCTION

The rapid pace of adoption of Web-enabled mobile handsets in worldwide markets has become an increasingly important issue for information systems professionals. A recent survey indicates that the number of global mobile Internet adopters is expected to reach nearly 600 million by 2008 (Ipsos-Insight, 2004; Probe Group, 2004), while the number of Internet-connected mobile phones will exceed the number of Internet-connected PCs by 2005 (*The Economist*, 2001). Such drastic convergence of the Internet and the mobile handset has been led by Asian and Scandinavian countries, where penetration has been especially meteoric. For example, roughly 70 million people in Japan, or 55% of the population, have signed up for mobile Internet access, in comparison to 12% in the United States (Faiola, 2004; Greenspan, 2003). Consequently, mobile phones or *Keitai* have been converted into devices for surfing the Internet, and by 2004 monthly mobile spending per consumer exceeded 35 euro.

Much of this success can be traced back to 1999, when NTT DoCoMo introduced the "i-mode" service. i-mode is a mobile service offering continuous Internet access based on packet-switching technology (Barnes & Huff, 2003). Through an i-mode handset, users can access a main micro-browser, which offers such typical services as e-mail, data search, instant messaging, Internet, and "i-menu." The "i-menu" acts as a mobile portal that leads to approximately 4,100 official and 50,000 unofficial sites (NTT DoCoMo 2003). Many such mobile portal sites can thus be considered as a pull-type advertising platform, where consumers can satisfy diverse information needs.

Several researchers have attempted to conceptualize the success of i-mode in comparison to WAP (Baldi & Thaung 2002) and in the light of the technology acceptance model (TAM) (Barnes & Huff 2003). Okazaki (2004) examined factors influencing consumer adoption of the i-mode pull-type advertising platform. However, there is a dearth of empirical research in this area, and especially in developing a model that captures the specific dimensions of mobile Internet adoption. In this respect, this study aims to propose a measurement scale of consumer perceptions of mobile portal sites.

The present study adopts, as its principal framework, the attitudinal model suggested by Dabholkar (1994). This includes "ease of use," "fun," and "performance" as important determinants of attitude. These are often referred to as "ease of use," "usefulness," and "enjoyment" in, for example, the TAM proposed by Davis (1986; Davis, Bagozzi, & Warshaw, 1989, 1992). The relevant literature suggests that dimensions similar to "ease of use" and "fun" are important antecedents of new technology adoption. For example, Shih (2004) and Szymanski and Hise (2000) found "perceived ease of use" and "convenient," respectively, as important antecedents of online behavior. Likewise, Moon and Kim (2001) found "perceived playfulness" to be a factor influencing WWW usage behavior, similar to the "fun" dimension. However, unlike earlier studies of m-commerce adoption, this study drops the third dimension of the TAM, "usefulness," in favor of "performance," because the former is appropriate only for tangible products, but not relevant for technology-based services (Dabholkar & Bagozzi, 2002). In contrast, "performance" represents a dimension that encompasses the reliability and accuracy of the technology-based service, as perceived by the consumer (Dabholkar, 1994). These three dimensions capture customer perceptions, which would initiate the attitude-intention-behavior causal chain (Davis, 1986).

BACKGROUND

Prior Theories on Technology Adoption

The technology acceptance model has been used to explain online user behavior (Featherman & Pavlous, 2002; Moon & Kim, 2001). Originally, TAM was based on Ajzen and Fishbein's (1980) theory of reasoned action (TRA), which is concerned with the determinants of consciously intended behaviors. TRA has been described as one of the most widely studied models in social psychology (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975). According to TRA, "a person's performance of a specified behavior is determined by his or her behavioral intention (BI) to perform the behavior, and BI is jointly determined by the person's attitude (A) and subjective norm (SN) concerning the behavior in question (Figure 1), with relative weights typically estimated by regression: $BI = A + SN$" (Davis et al., 1989). Here, BI refers to the degree of strength of one's intention to perform a specified behavior, while A is defined as an evaluative effect regarding performing the target behavior. SN is meant to be "the person's perception that most people who are important to

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Theory of reasoned action (TRA)

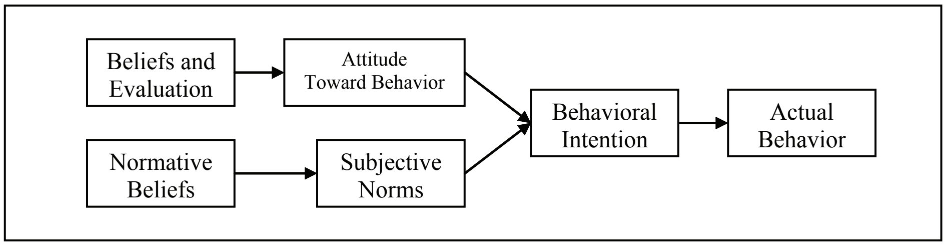
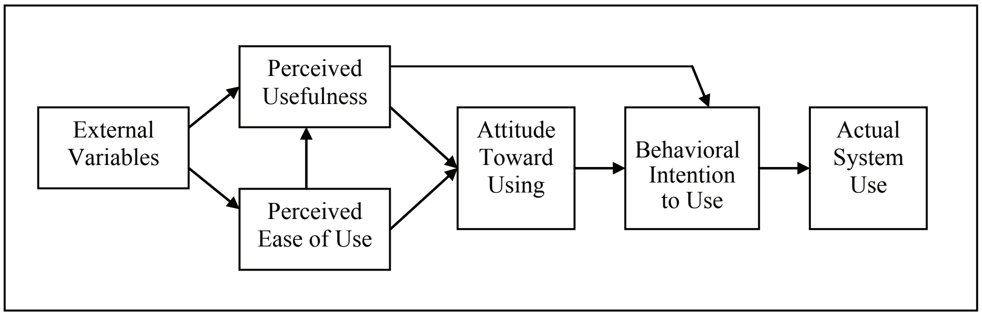


Figure 2. Technology acceptance model (TAM)



him think he should or should not perform the behavior in question" (Fishbein & Ajzen, 1975).

TAM extends TRA with attempts to explain the antecedents of computer-usage behavior. TAM comprises five fundamental salient beliefs: perceived ease of use, perceived usefulness, attitudes toward use, intention to use, and actual use. Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance," while perceived ease of use is "the degree to which a person believes that using a particular system would be free of effort" (Davis et al., 1989). Although they are not the only variables of interest in explaining user behavior, perceived ease of use and perceived usefulness have been proven empirically to be key determinants of behavior in a wide range of academic disciplines, such as the learning process of a computer language, evaluation of information reports, and adoption of alternative communication technologies, among others. However, TAM excludes the "influence of social and personal control factors on behavior" (Taylor & Todd, 1995). Consequently, Ajzen and Fishbein (1980) proposed another extension of TRA, the theory of planned behavior (TPB), to account for those conditions in which individuals may not have complete control over their own behavior (Taylor & Todd, 1995).

A key objective of TAM is "to assess the value of IT to an organization and to understand the determinants of that value" (Taylor & Todd, 1995). Hence, much IT research has aimed to enhance companies' effective IT resource management. However, TAM has been expanded to emerging new forms of IT, such as the wired as well as the wireless Internet.

In a pioneering study, Moon and Kim (2001) conducted empirical research on extending TAM to the World Wide Web context. They constructed an extension of TAM based on an individual's intrinsic motivation theory, and found that "perceived playfulness" had a positive effect on individuals' attitudes toward using the WWW and their behavioral intentions to use the WWW. Furthermore, TAM has been applied to explain the adoption of telecommunication technology, such as telework (Hun, Ku, & Chang, 2003), mobile devices (Kwon & Chidambaram, 2000), and m-commerce services (Pedersen & Ling, 2003). Generally, these studies suggest certain modifications of the original TAM in order to include social influence and behavioral control variables (Pedersen & Ling, 2003). In the following section, we explore the possible extension of TAM to m-commerce adoption.

Gender Differences in Mobile Internet Adoption

Gender has frequently been used as part of the social and the cultural meanings associated with developing marketing strategy via advertising messages, and in market segmentation strategy in particular, because it is easily: (1) identifiable, (2) accessible, (3) measurable, (4) responsive to marketing mix, (5) sufficiently large, and (6) profitable (Darley & Smith, 1995). However, although there has been much research on

Table 1. Rotated component matrix

		Component 1	Component 2	Component 3
Performance	Detailed	.695		
	Updated	.639		
	Intelligible	.619		
	Reliable	.449		
Ease of Use	Easy		.644	
	Killing Time		.589	
	Interesting		.536	
	Free		.440	
	Educational		.410	
Fun	Appealing			.642
	Helpful			.629
	Practical			.437
Total Variance		21.1	30.8	40.0
Eigenvalue		2.53	1.17	1.08

new technology adoption, little attention has been paid to gender differences in electronic communication. Yang and Lester (2005) argue that "research on gender and CMC has consistently demonstrated that gender inequalities define professional and scholarly electronic communication and that men are over-represented in electronic communities". This is considered a serious lacuna, since evidence has been found of important gender differences in human communication, including advertising (Wolin, 1999).

Our literature review found only one study that examined gender differences in online purchasing behavior. Yang and Lester (2005) conducted a series of studies on purchasing textbooks online at universities, and found that female students at an urban university tended to demonstrate fewer computer/Internet skills than male students, and that their level of skill was a more consistent predictor of purchasing textbooks online: the higher their level of skill, the more likely female students were to buy books online, and the effect of level of skill was greater for female than for male students.

To date, no gender studies of mobile Internet adoption have been reported. However, following Yang and Lester (2005), we may assume that, in learning and accessing wireless Internet with mobile handsets, female users may be less skillful than their male counterparts. For example, in terms of TAM, females may perceive more negatively ease of use, which is one of the essential determinants of attitude toward new technology adoption.

PROPOSED MODEL OF CONSUMER MOTIVATIONS

Although the specific motivations to use *wired* and *wireless* Internet must differ between individuals, the *overall* motivations of online information search may be similar for the two media. Thus, we adopted three primary motivations from prior research on wired Internet adoption: (1) performance, (2) ease of use, and (3) fun. First, Shih (2004) empirically examined online purchasing behavior, and found perceived usefulness to be the major determinant of behavioral intentions to use the Internet, while perceived ease of use is a secondary determinant. We adopt these concepts as performance and ease of use. It has been pointed out that the term *performance* is preferred to *usefulness*, in the case of intangible technology adoption. Second, Moon and Kim (2001) introduced an additional determinant of attitude formation, perceived playfulness or fun, to capture WWW usage behavior. Hence, we propose these three constructs as the principal drivers or motivations of enhanced mobile Internet usage. These constructs are essentially in line with Davis et al.'s (1989) TAM, which has frequently been used to explain and predict user adoption of a new information technology. Hence, our aim in this study is to examine whether there are any important differences between male and female mobile Internet users in terms of these constructs.

Table 2. Logistic regression results

Theoretical Constructs	Variables	Mobile Site		Internet		Satellite TV		Newspaper		WOM	
Performance	Detailed	-.117		-.005		.209		-.070		.005	
	Updated	.070		-.053		-.117		.202	*	-.107	
	Intelligible	.117		.073		.036		-.025		.042	
	Reliable	-.832	**	-.419	**	-.206		-.022		-.057	
Ease of Use	Easy	.253	**	.294	**	.241		.134		.297	***
	Killing Time	.311	**	.131		.255	*	-.285	**	-.345	**
	Interesting	.126		-.023		-.210		.098		.206	*
	Free	-.490	*	-.531	***	-.712	**	-.218		-.413	***
	Educational	-.042		-.021		.043		.137		-.318	**
Fun	Appealing	-.566		-.101		.021		-.226		.017	
	Helpful	-.019		.133		-.505		-.138		.518	**
	Practical	.103		.197		-.364	*	.189		.835	***

SURVEY METHOD

The survey was conducted via an online questionnaire that was made available in a popular commercial Web site in Japan. There were no restrictions on access, and the survey was open to the public audience. The questionnaire consists of a variety of questions, on general demographics, media usage, habits and spending, motives to use mobile Internet site, and so forth. As an incentive to complete the questionnaire, respondents were given an e-coupon as a reward for their participation. In total, 1,637 responses were obtained.

We assigned four adjectives for each of the three constructs: detailed, reliable, educational, and updated for performance; interesting, appealing, helpful, and killing time for fun; and easy, free, intelligible, and practical for ease of use. In order to identify the importance of each item, we used a dichotomous measure, asking whether respondents perceived a given adjective as describing his or her own perception of the mobile Internet site. For example, if they accessed a mobile Internet site because it seemed "reliable," they marked the answer "yes." In order to conduct statistical analysis, these dichotomous variables were converted into fictitious variables by assigning "1" to "yes" and "0" to "no."

RESULTS

With regard to the demographic composition by gender, the distribution of age and marital status differ little across gender; important differences can be observed in education and occupation. The proportion of people with bachelor or higher degrees is much greater in males than in females. On the other hand, females dominate junior college graduates. With regard to occupation, administrative, managerial, and professional workers are primarily male. A similar tendency can be seen in self-employed and skilled labor, although the magnitude is much less. There are more female workers in services.

To examine the dimensionality of the variables, we first conducted an exploratory factor analysis (EFA) with a principal component method. Although dichotomous variables are not ideal in EFA, fictitious variables are considered acceptable in this usage (Hair, Anderson, Tatham, & Black, 1998). The Varimax rotation was used, while a scree plot was carefully examined. Only variables with eigenvalue greater than 1 were retained. After several attempts using trial and error, we determined a three-factor solution to be the best, in which 12 proposed items were converged. However, as Table 1 shows, some of the items were classified into different constructs. Because of the exploratory nature of the study, we deemed this convergence to be reasonable and acceptable for the subsequent analysis.

Next, a logistic regression was performed with gender as a dependent variable and the importance (existence or absence) of adjective items as independent variables. It was possible to use binary data for both dependent and independent variables, because logistic regression does not require the normality assumption, as multiple regression does (Hair et al., 1998). However, because multicollinearity can seriously distort the results, a diagnostic was carried out via VIF and Tolerance values. Both values for each independent variable ranged from .80 to 1.23, showing no serious presence of multicollinearity.

The results of logistic regression are shown in Table 2. As clearly shown, ease of use plays an important role in separating male and female mobile Internet users. Chi-square tests reveal significant differences between male and female users in terms of easy, killing time, and free. Interestingly, female users are likely to perceive mobile Internet sites as an easy medium for killing time significantly more than their male counterparts. The opposite is true for free: male users essentially appreciate a mobile Internet site as a free information source. With regard to reliability in performance, male users tend to perceive this item more strongly than female users. Finally, logistic regression was also performed for different media, such as (wired) Internet, satellite TV, newspapers, and word of mouth (WOM). Despite the dangers of oversimplification, it seems that a mobile Internet site exhibits the combined effects of a wired Internet and word of mouth.

IMPLICATIONS

Our findings clearly show that there are important differences between male and female mobile users in terms of motivations to access mobile Internet sites. Specifically, female users are more prone to access a mobile Internet site for spare-time leisure and ease of use, while male users do so for free information. It should be noted that both genders perceive a mobile Internet site as a reliable, updated, and intelligible information source. In comparison with other media, a mobile Internet site is considered to be a combination of Internet and word of mouth. This makes sense because a mobile device is essentially and uniquely characterized as a personalized telecommunication medium, which is accessible only via a mobile telephone.

Given that Japanese mobile Internet services focus on information and entertainment (Okazaki, 2004), the findings of this study may provide useful implications for IT managers. That is, female users are more likely to appreciate a mobile Internet site for its entertainment value, while male users may seek more pragmatic results or outcomes, that is, reliable daily information. For example, typical male white-collar workers may seek daily financial news and replace newspapers with a mobile device as an information source. On the other hand, female users are attracted by more enjoyable usage, which provides an easy escape from daily routine. In this regard, IT managers should be aware of the importance of tailoring the content of mobile Internet according to gender-specific needs and wants.

Limitations

A few limitations should be recognized to make our findings more objective. First, our study examined only 12 adjective items with three proposed constructs. Future research should include a larger variety of items that may be related to consumers' perceptions associated with mobile Internet sites. Second, this study did not specify a type of "mobile Internet site." That is, our findings should be interpreted at most as general evaluations of mobile Internet adoption. Given a rapid proliferation of mobile Internet services, future research should specify the type of mobile Internet site and its benefits as a unit of analysis. Finally, the binary nature of data means that the construct reliability and validity were not established. Any future study should use a semantic differential scale, instead of a dichotomous scale, as a measure, and much effort should be made to improve the reliability indices.

REFERENCES

Ajzen, I. (1985). From intentions to actions: A theory of planned behaviour. In *Action control: From cognition to behaviour* (pp. 11-39). New York: Springer-Verlag.

Darley, W., & Smith, R. (1995). Gender differences in information processing strategies: An empirical test of the selectivity model in advertising response. *Journal of Advertising, 24*(1), 41-56.

Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Durlacher. (1999, November). *Mobile commerce report.* Retrieved from http://www.durlacher.com/fr-research-reps.htm

Featherman, M., & Pavlos, P. (2002). Predicting e-services adoption: A perceived risk facets perspective. *Proceedings of AMCIS 2002,* Dallas, TX.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, behaviour: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Hair, J. Jr., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Höflich, J., & Rössler, P. (2001). Mobile schriftliche Kommunikation oder: E-mail fürdas handy. *Medien & Kommunikationswissenschaft, 49,* 437-461. Cited by Pedersen & Ling (2003).

Hun, S., Ku, C., & Chang, C. (2003). Critical factors of WAP services adoption: An empirical study. *Electronic Commerce Research and Applications, 2*(1), 42-60.

Juniperresearch.com. (2004). Mobile data sales predicted to bolster operator revenues. *New Media Age,* (October 21), 6.

Kleijnen, M., Wetzels, M., & Ruyter, K. (2004). Consumer acceptance of wireless finance. *Journal of Financial Service Marketing, 8*(3), 206-217.

Lin, C. (1996). Looking back: The contribution of Blumler and Katz's 'Uses of mass communication' to communication research. *Journal of Broadcasting & Electronic Media, 40*(4), 574-581.

Moon, J., & Kim, Y. (2001). Extending the TAM for a World-Wide-Web context. *Information & Management, 38,* 217-230.

NTT DoCoMo. (2003, October 30). *I-mode subscribers surpass 40 million.* Retrieved January 2004 from http://www.nttdocomo.com/

Okazaki, S. (2004). How do Japanese consumers perceive wireless advertising? A multivariate analysis. *International Journal of Advertising, 23*(4), 429-454.

Pagani, M. (2004). Determinants of adoption of third generation mobile multimedia services. *Journal of Interactive Marketing, 18*(3), 46-59.

Pedersen, P., & Ling, R. (2003). Modifying adoption research for mobile Internet service adoption: Cross-disciplinary interactions. *Proceedings of the 36th IEEE Hawaii International Conference on System Sciences 2003,* (pp. 90-91).

Sadeh, N. (2002). *M-commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

Shih, H. (2004). Extended technology acceptance model of Internet utilization behaviour. *Information & Management, 41,* 719-729.

Taylor, S., & Todd, P. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144-176.

Wolin, L. (2003). Gender issues in advertising—An oversight synthesis of research: 1970-2002. *Journal of Advertising Research, 43,* 111-129.

Yang, B., & Lester, D. (2005). Sex differences in purchasing textbooks online. *Computers in Human Behaviour, 21,* 147-152.

KEY TERMS

i-mode: A broad range of Internet services for a monthly fee of approximately 3 Euro, including e-mail, transaction services (e.g., banking, trading, shopping, ticket reservations, etc.), infotainment services (e.g., news, weather, sports, games, music download, karaoke, etc.), and directory services (e.g., telephone directory, restaurant guide, city information, etc.), which offers more than 3,000 official sites accessible through the i-mode menu.

Mobile Commerce (M-Commerce): Any transaction with a monetary value that is conducted via a mobile telecommunications network. In a broader sense, it can be defined as the emerging set of applications and services people can access from their Internet-enabled mobile devices.

Mobile Portal: Typically, m-commerce takes place in a strategic platform called a "mobile portal," where third-generation (3G) mobile communication systems offer a high degree of commonality of worldwide roaming capability, support for a wide range of Internet and multimedia applications and services, and data rates in excess of 144 kbps. Examples of such mobile portals take many forms: NTT DoCoMo's i-mode portal, Nordea's WAP Solo portal, Webraska's SmartZone platform, among others. So far, Japan's i-mode portal has been asserted to be "the most successful and most comprehensive example of m-commerce today."

Technology Acceptance Model (TAM): Extends TRA with attempts to explain the antecedents of computer-usage behavior. TAM comprises five fundamental salient beliefs: perceived ease of use, perceived usefulness, attitudes toward use, intention to use, and actual use.

Theory of Reasoned Action (TRA): This model explains that a person's performance of a specified behavior is determined by his or her behavioral intention (BI) to perform the behavior, and BI is jointly determined by the person's attitude (A) and subjective norm (SN) concerning the behavior in question, with relative weights typically estimated by regression: BI = A + SN.

Uses and Gratifications Theory: A theory derived from media communication studies that focuses on individual users' needs or motivations of a particular medium. According to a recent study of mobile phone users, seven gratifications were identified: fashion/status, affection/sociability, relaxation, mobility, immediate access, instrumentality, and reassurance.

Handheld Computing and J2ME for Internet-Enabled Mobile Handheld Devices

Wen-Chen Hu
University of North Dakota, USA

Jyh-haw Yeh
Boise State University, USA

I-Lung Kao
IBM, USA

Yapin Zhong
Shandong Institute of Physical Education and Sport, China

INTRODUCTION

Mobile commerce or m-commerce is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of Internet-enabled mobile handheld devices (Hu, Lee, & Yeh, 2004). It is expected to be the next milestone after electronic commerce blossoming in the late-1990s. Internet-enabled mobile handheld devices are one of the core components of a mobile commerce system, making it possible for mobile users to directly interact with mobile commerce applications. Much of a mobile user's first impression of the application will be formed by his or her interaction with the device, therefore the success of mobile commerce applications is greatly dependent on how easy they are to use. However, programming for handheld devices is never an easy task not only because the programming languages and environments are significantly different from the traditional ones, but also because various languages and operating systems are used by handheld devices and none of them dominates.

This article gives a study of handheld computing, especially J2ME (Java 2 Platform, Micro Edition) programming, for mobile commerce. Various environments/languages are available for client-side handheld programming. Five of the most popular are (1) BREW, (2) J2ME, (3) Palm OS, (4) Symbian OS, and (v) Windows Mobile. They apply different approaches to accomplishing the development of mobile applications. Three themes of this article are:

1. Introduction of handheld computing, which includes server- and client- side computing.
2. Brief introductions of four kinds of client-side computing.
3. Detailed discussion of J2ME and J2ME programming.

Other important issues such as a handheld computing development cycle will also be discussed.

BACKGROUND

Handheld computing is a fairly new computing area and a formal definition of it is not found yet. Nevertheless, the authors define it as follows: Handheld computing is the programming for handheld devices such as smart cellular phones and PDAs (personal digital assistants). It consists of two kinds of programming: client- and server- side programming.

The definitions of client- and server- side computing are given as follows:

- **Client-Side Handheld Computing:** It is the programming for handheld devices and it does not need the support from server-side programs. Typical applications created by it include (1) address books, (2) video games, (3) note pads, and (4) to-do list.
- **Server-Side Handheld Computing:** It is the programming for wireless mobile handheld devices and it needs the support from server-side programs. Typical applications created by it include (1) instant messages, (2) mobile Web contents, (3) online video games, and (4) wireless telephony.

This article will focus on the client-side computing. The server-side computing is briefly given next.

Server-Side Handheld Computing

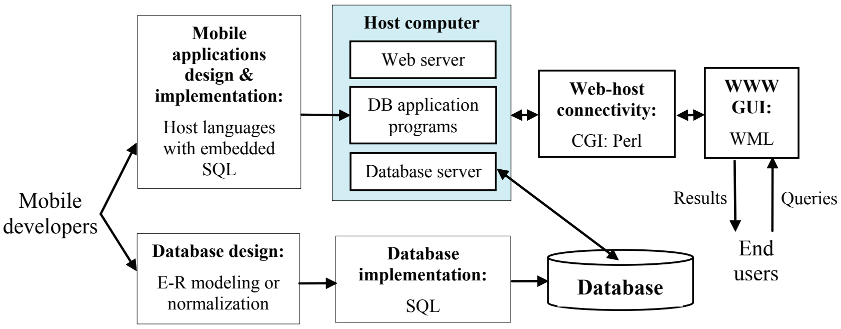
Most applications created by this kind of programming, such as instant messaging, require network programming such as TCP/IP programming, which will not be covered in this chapter. The most popular application of server-side handheld computing is database-driven mobile Web sites, whose structure is shown in Figure 1. A database-driven

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. A comparison among five handheld-computing languages/environments

	BREW	**J2ME**	**Palm OS**	**Symbian OS**	**Windows Mobile**
Creator	Qualcomm Inc.	Sun Microsystems Inc.	PalmSource Inc.	Symbian Ltd.	Microsoft Corp.
Language/ Environment	Environment	Language	Environment	Environment	Environment
Market Share (PDA) as of 2004	N/A	N/A	2nd	N/A	1st
Market Share (Smartphone) as of 2005	?	N/A	3rd	1st	2nd
Primary Host Language	C/C++	Java	C/C++	C/C++	C/C++
Target Devices	Phones	PDAs & phones	PDAs	Phones	PDAs & phones

Figure 1. A generalized system structure of a database-driven mobile Web site


Mobile applications design & implementation:
Host languages with embedded SQL
Host computer
Web server
DB application programs
Database server
Web-host connectivity:
CGI: Perl
WWW GUI:
WML
Results
Queries
End users
Mobile developers
Database design:
E-R modeling or normalization
Database implementation:
SQL
Database

mobile Web site is often implemented by using a three-tiered client/server architecture consisting of three layers:

1. **User Interface:** It runs on a handheld device (the client) and uses a standard graphical user interface (GUI).
2. **Functional Module:** This level actually processes data. It may consist of one or more separate modules running on a workstation or application server. This tier may be multi-tiered itself.
3. **Database Management System (DBMS):** A DBMS on a host computer stores the data required by the middle tier.

The three-tier design has many advantages over traditional two- or single- tier design, the chief one being: the added modularity makes it easier to modify or replace one tier without affecting the other tiers.

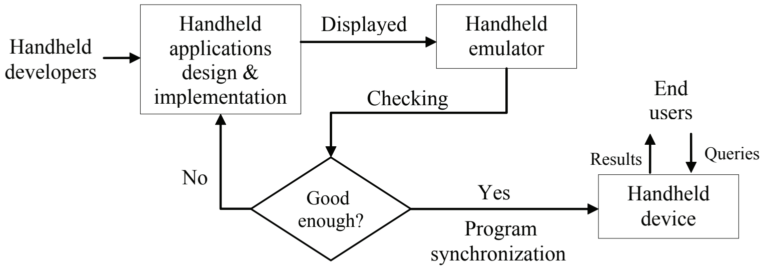
CLIENT-SIDE HANDHELD COMPUTING

Various environments/languages are available for client-side handheld programming. Five of the most popular are (1) BREW, (2) J2ME, (3) Palm OS, (4) Symbian OS, and (5) Windows Mobile. They apply different approaches to accomplishing the development of mobile applications. Figure 2 shows a generalized development cycle applied by them and Table 1 gives the comparison among the five languages/environments. The second half of this article is devoted to J2ME details and brief introductions of the other four are given in this section.

BREW (Binary Runtime Environment for Wireless)

BREW is an application development platform created by Qualcomm Inc. for CDMA-based mobile phones (Qualcomm Inc., 2003). CDMA is a digital wireless telephony transmission technique and its standards used for 2G mobile

Figure 2. A generalized client-side handheld computing development cycle



telephony are the IS-95 standards championed by Qualcomm. BREW is a complete, end-to-end solution for wireless applications development, device configuration, application distribution, and billing and payment. The complete BREW solution includes

- BREW SDK (software development kit) for application developers,
- BREW client software and porting tools for device manufacturers, and
- BREW distribution system (BDS) that is controlled and managed by operators—enabling them to easily get applications from developers to market and coordinate the billing and payment process.

Palm OS

Palm OS, developed by Palm Source Inc., is a fully ARM-native, 32-bit operating system running on handheld devices (PalmSource Inc., 2002). Palm OS runs on almost two out of every three PDAs. Its popularity can be attributed to its many advantages, such as its long battery life, support for a wide variety of wireless standards, and the abundant software available. The plain design of the Palm OS has resulted in a long battery life, approximately twice that of its rivals. It supports many important wireless standards, including Bluetooth and 802.11b local wireless and GSM, Mo-bitex, and CDMA wide-area wireless networks. Two major versions of Palm OS are currently under development:

- **Palm OS Garnet:** It is an enhanced version of Palm OS 5 and provides features such as dynamic input area, improved network communication, and support for a broad range of screen resolutions including QVGA.
- **Palm OS Cobalt:** It is Palm OS 6, which focuses on enabling faster and more efficient development of smartphones and integrated wireless (WiFi/Bluetooth) handhelds.

As of August 2005, no hardware products run Palm OS Cobalt and all devices use Palm OS Garnet. Likely as a result of Palm OS Cobalt's lack of adoption, PalmSource has shifted to developing Palm OS Cobalt's APIs on top of a Linux kernel.

Symbian OS

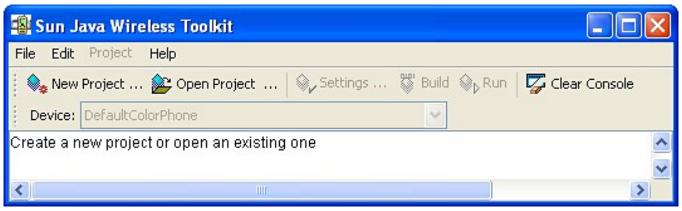
Symbian Ltd. is a software licensing company that develops and supplies the advanced, open, standard operating system—Symbian OS—for data-enabled mobile phones (Symbian Ltd., 2005). It is an independent, for-profit company whose mission is to establish Symbian OS as the world standard for mobile digital data systems, primarily for use in cellular telecoms. Symbian OS includes a multi-tasking multithreaded core, a user interface framework, data services enablers, application engines, integrated PIM functionality, and wireless communications. It is a descendant of EPOC, which is a range of operating systems developed by Psion for handheld devices.

Windows Mobile

Windows Mobile is a compact operating system for mobile devices based on the Microsoft Win32 API (Microsoft Corp., 2005). It is designed to be similar to desktop versions of Windows. In 1996, Microsoft launched Windows CE, a version of the Microsoft Windows operating system designed specially for a variety of embedded products, including handheld devices. However, it was not well received primarily because of battery-hungry hardware and limited functionality, possibly due to the way that Windows CE was adapted for handheld devices from other Microsoft 32-bit desktop operating systems. Windows Mobile includes three major kinds of software:

- **Pocket PCs:** Pocket PC enables you to store and retrieve e-mail, contacts, appointments, games, exchange text messages with MSN Messenger, browse the Web, and so on.
- **Smartphones:** Smartphone supplies functions of a mobile phone, but also integrates PDA-type functional-

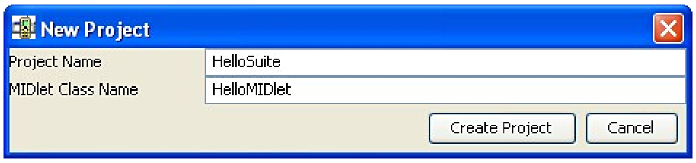
Figure 3. A screenshot of KToolbar after launching



Device: DefaultColorPhone

Create a new project or open an existing one

Figure 4. A screenshot of a pop-up window after clicking on the button New Project of KToolbar



Project Name: HelloSuite

MIDlet Class Name: HelloMIDlet

ity, such as e-mails, instant messages, music, and Web surfing, into a voice-centric handset.

- **Portable Media Centers:** Portable media centers let users take recorded TV programs, movies, home videos, music, and photos transferred from Microsoft Windows XP-based PC anywhere.

Windows Mobile-Based Pocket PCs

Pocket PCs were designed with better service for mobile users in mind and offers far more computing power than Windows CE. It provides scaled-down versions of many popular desktop applications, including Microsoft Outlook, Internet Explorer, Word, Excel, Windows Media Player, and others. It also includes three major kinds of software:

- **Pocket PC:** It puts the power of Windows software into a Pocket PC, giving you time to do more with the people and things that matter.
- **Pocket PC Phone Edition:** It combines all the standard functionality of a Windows Mobile-based Pocket PC with that of a feature-rich mobile phone.
- **Ruggedized Pocket PC:** It lets you do more of what matters to you even in the toughest user environments.

Windows Mobile-Based Smartphones

Windows Mobile-based smartphone integrates PDA-type functionality into a voice-centric handset comparable in size to today's mobile phones. It is designed for one-handed operation with keypad access to both voice and data features. The Smartphone is a Windows CE-based cellular phone. Like the Pocket PC, all Smartphones regardless of manufacturer share the same configuration of Windows CE. Also, Smartphones come bundled with a set of applications such as an address book, calendar, and e-mail program.

J2ME (JAVA 2 PLATFORM, MICRO EDITION)

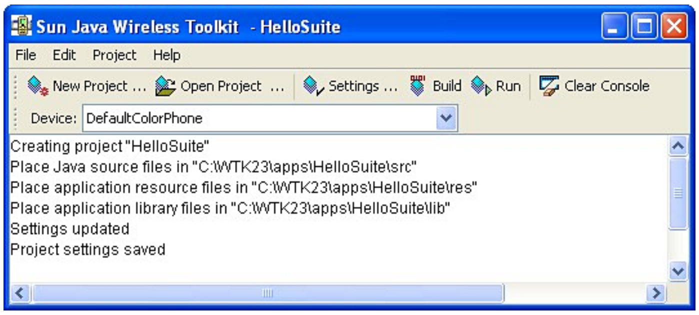
J2ME provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well as a broad range of embedded devices (Sun Microsystem Inc., 2002a). Like its counterparts for the enterprise (J2EE), desktop (J2SE) and smart card (Java Card) environments, J2ME includes Java virtual machines and a set of standard Java APIs defined through the Java Community Process, by expert groups whose members include device manufacturers, software vendors, and service providers.

J2ME Architecture

The J2ME architecture comprises a variety of configurations, profiles, and optional packages that implementers and developers can choose from, and combine to construct a complete Java runtime environment that closely fits the requirements of a particular range of devices and a target market. There are two sets of J2ME packages, which target different devices:

- **High-End Devices:** They include connected device configuration (CDC), foundation and personal profile.
- **Entry-Level Devices and Smart Phones:** They include connected limited device configuration (CLDC) and mobile information device profile (MIDP).

Figure 5. A screenshot of KToolbar after a project HelloSuite created



Device: DefaultColorPhone

```
Creating project "HelloSuite"
Place Java source files in "C:\WTK23\apps\HelloSuite\src"
Place application resource files in "C:\WTK23\apps\HelloSuite\res"
Place application library files in "C:\WTK23\apps\HelloSuite\lib"
Settings updated
Project settings saved
```

Figure 6. An example of an MIDlet program HelloMIDlet.java

C:\WTK23\apps\HelloSuite\src\HelloMIDlet.java

```
// This package defines MIDP applications and the interactions between
// the application and the environment in which the application runs.
import  javax.microedition.midlet.*;

// This package provides a set of features for user interfaces.
import  javax.microedition.lcdui.*;

public class HelloMIDlet  extends MIDlet  implements CommandListener {

  public void  startApp( ) {
    Display  display     = Display.getDisplay( this );
    Form     mainForm    = new  Form    ( “HelloMIDlet” );
    Ticker   ticker      = new  Ticker ( “Greeting, World” );
    Command  exitCommand = new  Command( “Exit”, Command.EXIT, 0 );

    mainForm.append            ( “\n\n          Hello, World!” );
    mainForm.setTicker         ( ticker );
    mainForm.addCommand        ( exitCommand );
    mainForm.setCommandListener( this );
    display.setCurrent         ( mainForm );
  }

  public void  pauseApp ( ) { }

  public void  destroyApp( boolean unconditional ) {
    notifyDestroyed( );
  }

  public void  commandAction( Command c, Displayable s ) {
    if ( c.getCommandType( ) == Command.EXIT )
      notifyDestroyed( );
  }
}
```

Configurations comprise a virtual machine and a minimal set of class libraries and they provide the base functionality for a particular range of devices that share similar characteristics, such as network connectivity and memory footprint. Profiles provide a complete runtime environment for a specific device category.

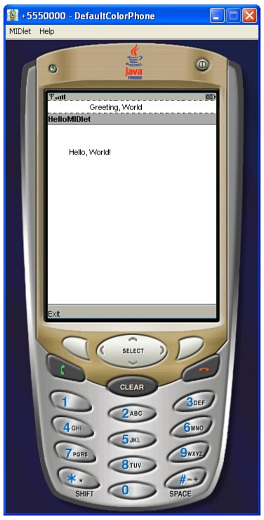
J2ME Programming

This sub-section gives an example of J2ME programming (Sun Microsystem Inc., 2004). Other client-side handheld programming is similar to this. Figure 3 shows the Sun Java Wireless Toolkit©, which is a toolbox for developing wireless applications that are based on J2ME's CLDC and MIDP. The

Table 2. Mobile Information Device Profile (MIDP) package list

Package	Classes and Descriptions
User Interface	**`javax.microedition.lcdui`**: The UI API provides a set of features for implementation of user interfaces for MIDP applications.
	`javax.microedition.lcdui.game`: The Game API package provides a series of classes that enable the development of rich gaming content for wireless devices.
Persistence	**`javax.microedition.rms`**: It provides a mechanism for MIDlets to persistently store data and later retrieve it.
Application Lifecycle	**`javax.microedition.midlet`**: The MIDlet package defines MIDP applications and the interactions between the application and the environment in which the application runs.
Networking	**`javax.microedition.io`**: The MID Profile includes networking support based on the `Generic Connection` framework from the *Connected, Limited Device Configuration*.
Audio	**`javax.microedition.media`**: The MIDP 2.0 Media API is a directly compatible building block of the Mobile Media API (JSR-135) specification.
	`javax.microedition.media.control`: This package defines the specific Control types that can be used with a `Player`.
Public Key	**`javax.microedition.pki`**: Certificates are used to authenticate information for secure Connections.
Core	**`java.io`**: Provides classes for input and output through data streams.
	`java.lang`: MID Profile Language Classes included from Java 2 Standard Edition.
	`java.util`: MID Profile Utility Classes included from Java 2 Standard Edition.

Figure 7. A screenshot of an emulator displaying the execution results of HelloSuite



Greeting, World

Hello, World!

toolkit includes the emulation environments, performance optimization and tuning features, documentation, and examples that developers need to bring efficient and successful wireless applications to market quickly. The following steps show how to develop an MIDP application, a simple "Hello, World!" program, under Microsoft Windows XP:

1. Download Sun Java Wireless Toolkit 2.3 Beta, which includes a set of tools and utilities and an emulator for creating Java applications that run on handheld devices, at http://java.sun.com/products/sjwtoolkit/download-2_3.html .
2. Run MIDlet, an MIDP application, development environment KToolbar as shown in Figure 3 by selecting the following Windows commands:

   Start ► All Programs ► Sun Java Wireless Toolkit 2.3 Beta ► KToolbar

3. Create a new project by giving a project name such as HelloSuite and a class name such as HelloMIDlet as shown in Figure 4. After the project HelloSuite is created, the KToolbar will display the message shown in Figure 5, which tells where to put the Java source files, application resource files, and application library files.
4. Create a J2ME source program and put it in the directory C:\WTK23\apps\HelloSuite\src\. Figure 6 gives a J2ME example, which displays the text "Hello, World!" and a ticker with a message "Greeting, world."
5. Build the project by clicking on the `Build` button. The `Build` includes compilation and pre-verifying.
6. Run the project by clicking on the `Run` button. An emulator will be popped up and displays the execution results of the built project. For example, Figure 7 shows an emulator displays the execution results of HelloSuite.
7. Upload the application to handheld devices by using USB cables, infrared ports, or Bluetooth wireless technology.

Mobile Information Device Profile (MIDP) Packages

Table 2 shows the packages provided by the MIDP (Sun Microsystem Inc., 2002b). The packages javax.* are the extensions to standard Java packages. They are not included in the JDK or JRE. They must be downloaded separately.

FUTURE TRENDS

A number of mobile operating systems with small footprints and reduced storage capacity have emerged to support the computing-related functions of mobile devices. For example, Rearch In Motion Ltd.'s BlackBerry 8700 smartphone uses RIM OS and provides Web access, as well as wireless voice, address book, and appointment applications (Research In Motion Ltd., 2005). Because the handheld device is small and has limited power and memory, the mobile OSs' requirements are significantly less than those of desktop OSs. Although a wide range of mobile handheld devices are available in the market, the operating systems, the hub of the devices, are dominated by just few major organizations. The following two lists show the operating systems used in the top brands of smart cellular phones and PDAs in descending order of market share:

- **Smart Cellular Phones:** Symbian OS, Microsoft Smartphone, Palm OS, Linux, and RIM OS (Symbian Ltd., n.d.).
- **PDAs:** Microsoft Pocket PC, Palm OS, RIM OS, and Linux (WindowsForDevices, 2004).

The market share is changing frequently and claims concerning the share vary enormously. It is almost impossible to predict which will be the ultimate winner in the battle of mobile operating systems.

CONCLUSION

Mobile commerce is a coming milestone after electronic commerce blossoming in the late-1990s. The success of mobile commerce applications is greatly dependent on handheld devices, by which mobile users perform the mobile transactions. Handheld computing is defined as the programming for handheld devices such as smart cellular phones and PDAs. It consists of two kinds of programming: client- and server- side programming. Various environments/languages are available for client-side handheld programming. Five of the most popular are

1. **BREW:** It is created by Qualcomm Inc. for CDMA-based smartphones.
2. **J2ME:** J2ME is an edition of the Java platform that is targeted at small, standalone or connectable consumer and embedded devices.
3. **Palm OS:** It is a fully ARM-native, 32-bit operating system running on handheld devices.
4. **Symbian OS:** Symbian OS is an industry standard operating system for smartphones, a joint venture originally set up by Ericsson, Nokia, and Psion.
5. **Windows Mobile:** Windows Mobile is a compact operating system for handheld devices based on the Microsoft Win32 API. It is a small version of Windows, and features ma ny "pocket" versions of popular Microsoft applications, such as Pocket Word, Excel, Access, PowerPoint, and Internet Explorer.

They apply different approaches to accomplishing the development of handheld applications and it is almost impossible to predict which approaches will dominate the client-side handheld computing in the future, as the Windows to desktop PCs.

REFERENCES

Hu, W.-C., Lee, C.-W., & Yeh, J.-H. (2004). Mobile commerce systems. In Shi Nansi (Ed.), *Mobile Commerce Applications* (pp. 1-23). Hershey, PA: Idea Group Publishing.

Microsoft Corp. (2005). *What's new for developers in Windows Mobile 5.0?* Retrieved August 29, 2005, from http://msdn.microsoft.com/mobility/windowsmobile/howto/documentation/default.aspx?pull=/library/en-us/dnppcgen/html/whatsnew_wm5.asp

PalmSource Inc. (2002). *Why PalmOS?* Retrieved June 23, 2005, from http://www.palmsource.com/palmos/Advantage/index_files/v3_document.htm

Qualcomm Inc. (2003). *BREW and J2ME: A complete wireless solution for operators committed to Java*. Retrieved February 12, 2005, from http://brew.qualcomm.com/brew/en/img/about/pdf/brew_j2me.pdf

Research In Motion Ltd. (2005). *BlackBerry application control: An overview for application developers*. Retrieved January 05, 2006, from http://www.blackberry.com/knowledgecenterpublic/livelink.exe/fetch/2000/7979/1181821/832210/BlackBerry_Application_Control_Overview_for_Developers.pdf?nodeid=1106734&vernum=0

Sun Microsystem Inc. (2002a). *Java 2 Platform, Micro Edition*. Retrieved January 12, 2006, from http://java.sun.com/j2me/docs/j2me-ds.pdf

Sun Microsystem Inc. (2002b). *Mobile information device profile specification 2.0*. Retrieved October 25, 2005, from http://jcp.org/aboutJava/communityprocess/final/jsr118/

Sun Microsystem Inc. (2004). *J2ME Wireless Toolkit 2.2: User's guide*. Retrieved October 21, 2005, from http://java.sun.com/j2me/docs/wtk2.2/docs/UserGuide.pdf

Symbian Ltd. (2005). *Symbain OS Version 9.2*. Retrieved December 20, 2005, from http://www.symbian.com/technology/symbianOSv9.2_ds_0905.pdf

Symbian Ltd. (n.d.). *Symbian fast facts*. Retrieved January 26, 2005, from http://www.symbian.com/about/fastfaqs.html

Wilson, J. (2005). *What's new for developers in Windows Mobile 5.0*. Retrieved January 14, 2006, from http://msdn.microsoft.com/smartclient/default.aspx?pull=/library/en-us/dnppcgen/html/whatsnew_wm5.asp&print=true

WindowsForDevices.com. (2004). *Windows CE zooms past Palm*. Retrieved August 23, 2005, form http://www.windowsfordevices.com/news/NS6887329036.html

KEY TERMS

Binary Runtime Environment for Wireless (BREW): BREW is an application development platform created by Qualcomm Inc. for CDMA-based mobile phones.

Client-Side Handheld Programming: It is the programming for handheld devices and it does not need the supports from server-side programs. Typical applications created by it include (1) address books, (2) video games, (3) note pads, and (4) to-do list.

Handheld Computing: It is the programming for handheld devices such as smart cellular phones and PDAs (Personal Digital Assistants). It consists of two kinds of programming: client- and server-side programming.

Java 2 Platform, Micro Edition (J2ME): J2ME provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well as a broad range of embedded devices.

Palm OS: Palm OS, developed by Palm Source Inc., is a fully ARM-native, 32-bit operating system running on handheld devices.

Server-Side Handheld Programming: It is the programming for wireless mobile handheld devices and it needs the supports from server-side programs. Typical applications created by it include (1) instant messages, (2) mobile Web contents, (3) online video games, and (4) wireless telephony.

Symbian OS: Symbian Ltd. is a software licensing company that develops and supplies the advanced, open, standard operating system—Symbian OS—for data-enabled mobile phones.

Windows Mobile: Windows Mobile is a compact operating system for mobile devices based on the Microsoft Win32 API. It is designed to be similar to desktop versions of Windows.

An Infrastructural Perspective on U-Commerce

Stephen Keegan
University College Dublin, Ireland

Caroline Byrne
Institute of Technology Carlow, Ireland

Peter O'Hare
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

INTRODUCTION

In modern mobile-equipped businesses, the scales of economics sway between increasing economic returns and flawlessly decreasing expenditures while providing a worthwhile service for their customer base. Early mobile computing adopters realized that the scales of economic solvency weighed in favor of businesses that seamlessly delivered and managed customer expectations. This is only feasible if all frontline staff are endowed with relevant technological advances and educated appropriately in their usage. Timely, adequate responses to customer requests results in retaining satisfied customers and an expanding customer base. Efficient use of mobile advances can reduce mundane office tasks by preventing replication of work through data transfer between mobile devices and workstations. These streamlined tasks can often tilt the scales favorably for a struggling company.

Mobile computing encourages technological advances at the company's cutting edge while supporting its employees' daily duties by optimizing tasks. This is achieved via various handheld devices, each operating daily as unique satellite data stations, wirelessly updating the central company computer system. Another recent phenomenon is that of astute consumers comparing and contrasting products and prices prior to purchase via the internet. Mobile computing allows us the luxury of comparison from our closest physical retail outlet. As we physically view the product desired, our mobile enabled handheld device can navigate the Internet for comparable products at more competitive prices, thus allowing us the power to purchase under the canopy of an informed choice.

We define u-commerce as "the use of ubiquitous networks to support personalized and uninterrupted communications and transactions between an organization and its various stakeholders to provide a level of value over, above, and beyond traditional commerce" (Junglas & Watson 2003). U-commerce encompasses concepts that are ubiquitous, universal, unique, and unison. We take this opportunity to explore each of these in some depth.

Ubiquitous

Computers are already ubiquitous in our society. With continually decreasing hardware costs, relentless miniaturization, and the adoption of high-speed networks, this trend is likely to continue. Modern automobiles already contain dozens of microprocessors, while the unabated popularity of third-generation mobile phones means that mobile computing is now within reach of people in their daily lives.

Universal

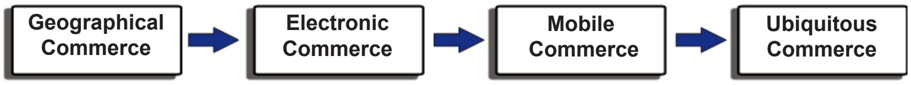
The utility factor of u-commerce-enabling accessories like laptops, mobile phones, and PDAs has been limited by the fact that they are often not universally usable. Perhaps the most well-known instance of this type of incompatibility lies within the domain of mobile phones. People traveling between Europe and the United States often find that their European (GSM) phones operating at the 900 MHz and/or 1800 MHz frequencies are incompatible with those in the United States (CDMA), which typically operate at a frequency of 1900 MHz.

Unique

Many current retail delivery systems fail to exploit the unique characteristics of each individual user. Within u-commerce we envisage a model whereby users interact with information and services based upon the context at that point in time. Here context can entail such factors as temporal informa-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. The development of commerce



tion (e.g., What time of day is it?), user preferences (Does the user like ice-cream?), location data (e.g., How far is the user from our shop?), or user profile data (e.g., Is this user a female tourist?).

Unison

U-commerce relies on unison between all electronic data and equipment relevant in the user's life. Appropriate data such as profile information, product preferences, and financial data is securely shared in a distributed fashion and is readily retrievable at the appropriate time. Unison delivers the integration of various communication systems so there is a single interface or connection point.

BACKGROUND

Over the past decade the emergence of new electronic mobile and communications technologies has driven the way we conduct our business. Traditionally, commerce was geographic, with consumers seeking out and physically purchasing a product or service. The rapid deployment and ready accessibility of the Internet led to the dawn of *electronic commerce* (e-commerce). E-commerce enabled consumers to purchase products and service electronically via the Internet (http://www.ebay.com, http://www.amazon.com).

The development and widespread deployment of wireless technologies has ensured that mobile computing is spawning a dominant new culture (Rheingold, 2002). The mobile culture has gripped modern society with people regularly using their cellular phones, PDAs, MP3 players, and digital cameras. The development of new wireless hardware, software, and services is now occurring at an exponential rate. As a result m-commerce and u- commerce applications and services must be developed if they wish to evolve with the available technology.

M-commerce and u-commerce have significant differences from the geographical and electronic commerce which preceded them. Mobile devices impose a number of constraints upon business and service providers, including: smaller screen sizes, reduced interface interactivity, shorter battery life, and a restricted computational power. These restrictions have direct implications upon the mobile consumers, with users being less tolerant of irrelevant information and as a consequence having a shorter attention span. M-commerce and u-commerce business and service providers must address these restrictions and resolve them in creative and intelligent ways.

STATE OF THE ART

Shoppers today face a bewildering array of choices, whether they are shopping online or in the real world. To help shoppers cope with all of these choices, online merchants have deployed recommender systems that guide people toward products they are more likely to find interesting (Sarwar, Karypis, Konstan, & Reidl, 2001). Many of these online recommender systems operate by suggesting products that complement products you have purchased in the past. Others suggest products that complement those you have in your shopping cart at checkout time. If you have ever bought a book at Amazon.com or browsed musical listings at yahoo.com, you may have used a recommender system. Some of these systems, though ingenious, can prove to be of limited utility when applied to a mobile scenario. Dynamic pricing, mobile users, and limited hardware capabilities mean that new approaches are imperative.

Movielens Unplugged (Miller, Albert, Lam, Konstan, & Riedl, 2003) attempted to transpose the usability of the Movielens project to a selection of mobile devices. Particular emphasis was placed on developing a user interface that was capable of supporting multiple front ends and multiple devices. A set of generalized design principles was derived during a user trial. MobyRek (Ricci, Nguyen, & Cavada, 2004) is an on-tour recommendation system that becomes operational when a mobile traveler requests MobyRek to find some interesting travel products and ends when the traveler either selects a product or quits the session. It evolves in cycles, and in each cycle a set of recommended products is shown to the user. The recommendation process that it employs consists of four logical components: initialization, interaction, adaptation, and retainment.

Mobitip (Rudström, Svensson, Cöster, & Höök, 2004) is a mobile recommender system that allows people to create, rate, and share information using short-range Bluetooth communication, while occasionally synchronizing with a central server. It is argued that a real-time distribution schema of user profile data in impractical. The proposed solution involves storing a user's profile on the mobile device together, with a ranked list of predictions from a central server computed the last time the user docked. These predictions are recalculated as and when new data becomes available.

Of particular interest to us are the several companies who are experimenting with using Bluetooth to deliver the personal shopping assistant vision. WideRay (http://www.wideray.com/) has placed several of its BlueRay kiosks in selected music outlets, video shops, and theatres at a number of locations in Europe and the United States. When customers come within 10 meters of the kiosks, they receive a text message asking if they are interested in getting more information about various items the store is selling, perhaps music, ring tones, videos, or games. If customers are interested, they can go to the kiosk and choose what to download. Moonstorm (http://www.moonstorm.com/) recently released software called Cellfire that can automatically download coupons to mobile phones for stores in the customer's area. Customers do not receive any intrusive text messages, but must click on the Cellfire icon on their phones to examine and use the coupons.

A recent project, eNcentive, (Ratsimor, Finin, Joshi, & Yesha, 2003), is an infrastructure that facilitates peer-to-peer electronic commerce in the mobile environment. The system functions by aggressively broadcasting coupons, advertisements, and promotions through a geographical region populated by businesses (i.e., restaurants, dry cleaners, etc.) onto users' PDAs.

iGrocer (Shekar, Nair, & Helal, 2003) is a smart grocery shopping assistant that is hosted on a smart phone that comes with a bar code scanner. iGrocer helps users create and maintain weekly grocery shopping lists. Another feature of iGrocer is the nutrition indicator, which recommends foods based on a compatibility check between the user's health profile and the nutritional value of the food. The device can also act as a 'trusted checkout' and thereby act on behalf of the store and the customer.

PARTICIPANTS

There are some prominent challenges implicit in delivering the u-commerce vision. Some of these may be familiar to the reader in that they are equally applicable to the e-commerce domain, while others are of particular pertinence in the domain of u-commerce. The ability to provide users with simple, convenient, and trusted means of purchasing goods and services is paramount to the realization of u-commerce. There are a number of different participants upon whom this depends.

The Consumer

The consumer embodies the demand side of all u-commerce product and service acquisition. It is the consumer who identifies a need at a particular point and seeks to consolidate that need in a convenient fashion as specified in the consumer buying behavior model (Guttman, Moukas, & Maes, 1998). We assume that the u-commerce consumer is equipped with a networked device or set of devices to support the timely acquisition of products and services. The terms *user* and *consumer* are interchangeable within the scope of the remainder of this article.

The Provisor

The provisor is analogous to a real-world retailer—that is, a *bricks-and-mortar store.* In the context of u-commerce, however, the provisor may or may not have a real-world presence. Each provisor competes for consumers on the basis of a dynamic set of information garnered from a range of sources. The consumer's personal profile may be stored on the consumer's device, while an annotated record of that consumer's past purchasing behavior may be retrieved from a different location at the behest of the provisor (assuming authorization from the consumer). This composite set of data is utilized by the provisor in making appropriate product offerings to the consumer.

The Mediator

The mediator acts as a conduit between buyer (consumer) and seller (provisor). Its principal role is to ensure that trade occurs between all parties in an efficient and fair manner. A secondary role is to maintain and coordinate access to additional infrastructure required for u-commerce and not directly supplied by the consumer or the provisor. The most obvious example of this kind of infrastructure is a data network; this class of participant includes fixed and mobile network operators.

CHALLENGES

After considering the requirements of these three participants, a series of challenges can be identified with relation to the implementation of a supportive u-commerce architecture. We can classify these challenges in accordance with these participants as follows.

Infrastructural Considerations

Since commercial transitions within the system are completely data-flow based: the rate at which information can be transferred around the system in an efficient and timely manner is of prime importance. This rate is dependent upon the capabilities of the underlying infrastructure. It is envisaged that u-commerce data transfer requirements will accelerate, with a growing number of consumers demanding an increasingly *immersive experience* with multimedia aspects. The continued evolution of mobile data network infrastructures can support this growth.

Hardware Constraints

Challenges such as limited memory and processing power, although of particular concern to the consumer, are considerable design factors for both the provisor and infrastructural operator. A multitude of (often heterogeneous) hardware modules are required to operate in tandem to deliver commercial services. Bluetooth hotspots, GPS units, and electronic compasses are examples of this type of module.

Human-Computer Interface Issues

As with hardware considerations, *human-computer interface* (HCI) issues are most relevant when discussing the interface requirements and capabilities of the consumer's portable (and therefore diminutive) hardware—for example, mobile phone, PDA, personal GPS unit. In a mobile device environment, more than any other platform, each layer of the application architecture must be carefully considered and prioritized to maximize the device's physical capacity.

Standards and Interoperability

The ability of different, sometimes competing, parties to buy and sell goods is dependent upon a set of well-defined interoperability metrics. A number of initiatives have emerged to counter this problem. The incorporation of XML (http://www.w3.org/XML/) and markup-derived ontologies is seen as a suitable mechanism to deliver fair and standardized trading channels. Foremost of these is the UNSPSC (http://www.unspsc.org/), an extendible system of 18,000 terms to classify both products and services jointly developed by the United Nations Development Program (UNDP) and Dun & Bradstreet Corporation (D&B) in 1998.

Cultural Aspects

The intrepidation sometimes observed in consumers when adopting new commerce channels is perhaps understandable. The level of this reluctance is not universal. The modest uptake in the recent and much-heralded introduction of *i-mode* services on European networks, when contrasted with the blistering growth of i-mode in recent years in its home market of Japan only, serves to reinforce the importance of cultural factors. The role of cultural acceptance is often underestimated by proponents of u-commerce who may be predisposed with a technical bias.

Security Concerns

In the late 1990s the potential of online shopping was seen as being underdeveloped. A primary factor was the well-founded concerns of potential online shoppers over security. Security on the seemingly anarchic structure of the World Wide Web is only as strong as its weakest link. Some important lessons can be drawn from this. Firstly, all facets of a u-commerce architecture must be secure, and perhaps more importantly the system must not be just secure, but must be seen as being secure.

Legal/Regulatory Issues

A major challenge for u-commerce is that legal and regulatory stipulations in some markets around the world sometimes prevent network operators from brokering the sale of non-communication goods or services without the legal status of a bank or financial institution. To overcome these restrictions, there needs to be close cooperation between:

- banks, as they are the trusted intermediaries between consumers and provisor;
- credit card issuers, with their global coverage and extensive experience;
- network operators, who maintain an established subscriber base and who hold a central position in the communication value chain; and
- provisors.

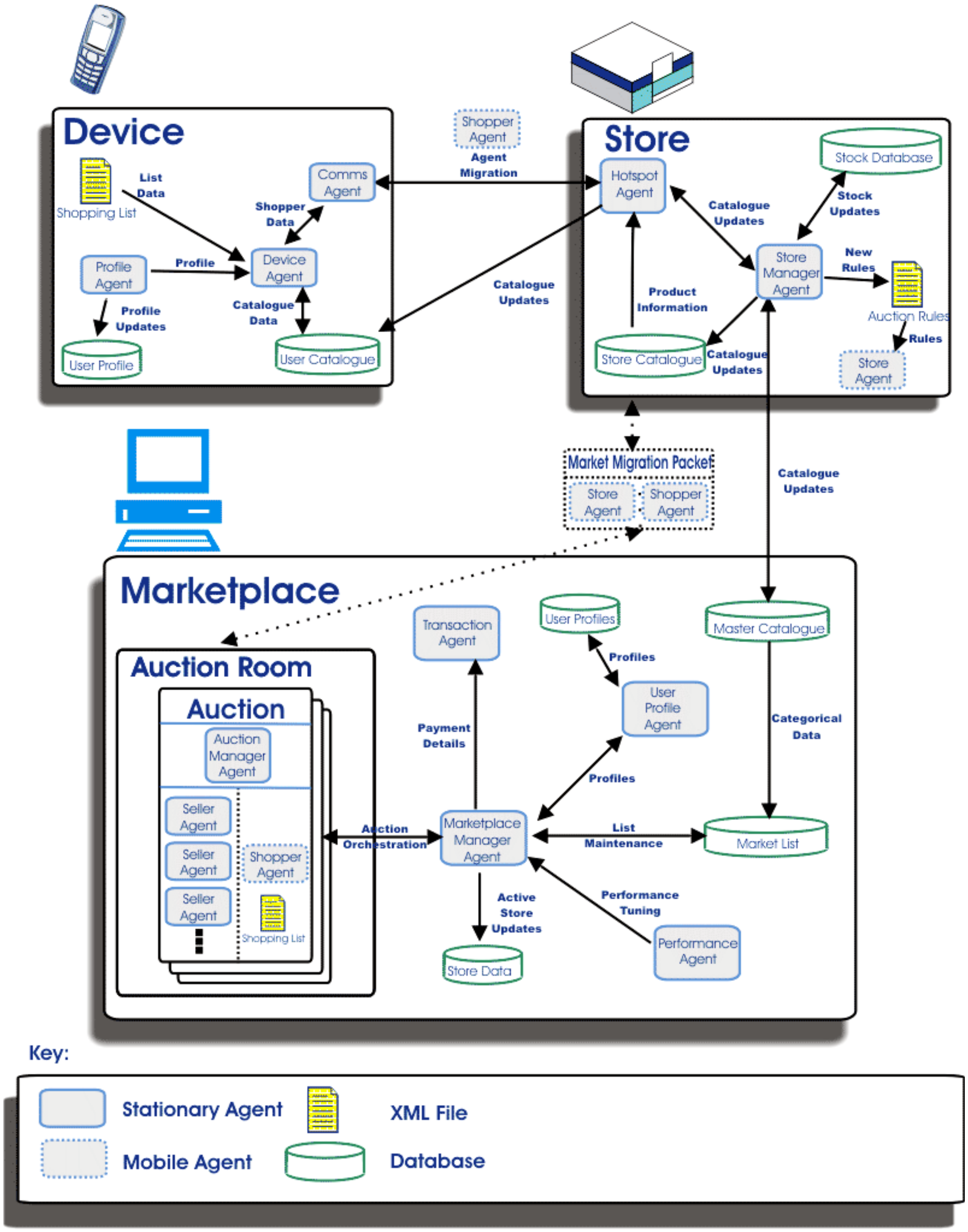
Viable and Fair Business Model

The rise to prominence of eBay and Amazon, among others, is partly due to the utilization of appropriate business models. Successful models recognize the need to ensure that all participants are rewarded. This need must be balanced with an assurance that the cost of airtime must not be prohibitive compared with the amount of the transaction. Revenue-sharing models must provide win/win resolutions for all parties, with each participant receiving a reward, however small, for all transactions in which they played a part.

Suitable Payment Solutions

A u-commerce payment system must be capable of integrating different provisors, financial institutions, as well as other payment systems. The payment solution must comply with existing legal regulations and be flexible enough to be *localized* in accordance with different practices around the world, and must be capable of meeting any future legal requirements.

Figure 2. Easishop system overview


Device
Shopping List
List Data
Comms Agent
Shopper Data
Profile Agent
Profile
Device Agent
Profile Updates
Catalogue Data
User Profile
User Catalogue
Shopper Agent
Agent Migration
Catalogue Updates
Store
Hotspot Agent
Stock Database
Catalogue Updates
Stock Updates
Store Manager Agent
New Rules
Product Information
Auction Rules
Rules
Store Catalogue
Catalogue Updates
Store Agent
Market Migration Packet
Store Agent
Shopper Agent
Catalogue Updates
Marketplace
Transaction Agent
User Profiles
Master Catalogue
Profiles
Auction Room
Auction
Auction Manager Agent
User Profile Agent
Payment Details
Categorical Data
Profiles
Seller Agent
Seller Agent
Shopper Agent
Seller Agent
Shopping List
Auction Orchestration
Marketplace Manager Agent
List Maintenance
Market List
Active Store Updates
Performance Tuning
Store Data
Performance Agent
Key:
Stationary Agent
XML File
Mobile Agent
Database

IMPLEMENTATION

An example of a typical u-commerce framework is Easishop (Keegan & O'Hare, 2004). A three-tiered architecture, the system is composed of a *marketplace* (mediator), a set of competing, independent *stores* (provisors), as well as a set of mobile *device* users (consumers). This architecture is represented in Figure 2.

Easishop has been implemented both on an archetypical mobile phone and PDA, namely the SonyEricsson P910i and HP IPAQ 3870 respectively. All software has been implemented in Java, with the J2ME variant being deployed on the mobile devices and the standard edition J2SE V.5 used on the Easishop network nodes. Over the Bluetooth connection, the serial port profile (SPP) is used while standard IP is employed between shop nodes. The shop nodes themselves have been implemented on standard workstations connected via Ethernet. All GUI elements have been realized using the Thinlet (http://www.thinlet.org/) toolkit.

The notion of agency is fundamental to Easishop. All agents have been designed and implemented in Agent Factory (Collier, O'Hare, Lowen, & Rooney, 2003). The resultant agents enable an effective mechanism for delivering u-commerce interoperation. Specifically, Agent Factory supports the creation of a type of mobile agent that is autonomous, situated, socially able, intentional, rational, and mobile. The reasoning mechanism used by such agents conforms to a belief-desire-intention (BDI) (Rao & Georgeff, 1996) architecture.

CONCLUSION

U-commerce virtually envelops consumers with a bewildering array of telecommunication extravaganzas, empowering all equally with opportunities to have their business/social/entertainment desires instantly fulfilled. The enhanced security features that are penetrating the mobile industry provide society with a panacea regarding any previous security concerns. The mobile moguls are constantly diversifying in order to retain their market share, and offering their customers the facility to buy/sell/gamble as and when they desire is an opportunity too lucrative to ignore.

Several hardware/software manufacturers are responding to market and consumer demands with a variety of devices and applications that facilitate consumers' frenetic lifestyles by assisting with grocery/retail shopping. Consumers will continue to benefit from technological advances as the costs of mobile hardware and per second billing steadily decrease, and our suppliers dazzle us with applications that enhance our daily interactions.

This wave of telecommunications forces us to re-engineer our beliefs and perceptions on what constitutes mobile usage in today's culture and embrace the possibility of having what we want, when we want it. Undoubtedly, providers have numerous hardware or memory challenges ahead and will fervently endeavor to absorb these issues and provide pertinent solutions. Competition increases productivity, and ultimately it will be consumers who benefit from the current economic race to the u-commerce summit.

REFERENCES

Brody, A.B., & Gottsman, E.J. (1999). Pocket BargainFinder: A handheld device for augmented commerce. *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing (HUC'99)*, Karlsruhe, Germany.

Collier, R. W., O'Hare, G. M. P., Lowen, T. D., & Rooney, C. F. B. (2003). Beyond prototyping in the factory of agents. *Proceedings of the 3rd International Central & Eastern European Conference on Multi-Agent Systems (CEEMAS 2003)*, Prague, Czech Republic.

Guttman, R. H., Moukas, A. G., & Maes, P. (1998). Agents as mediators in electronic commerce. *Electronic Markets, 8*(1), 22-27.

Junglas, I. A., & Watson, R. T. (2003). U-commerce: A conceptual extension of e- and m-commerce. *Proceedings of the International Conference on Information Systems,* Seattle, WA.

Keegan, S., & O'Hare, G. M. P. (2004). Easishop—Agent-based cross merchant product comparison shopping for the mobile user. *Proceedings of the 1st International Conference on Information & Communication Technologies: From Theory to Applications (ICTTA '04)*, Damascus, Syria.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003, January). MovieLens unplugged: Experiences with an occasionally connected recommender system. *Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI '03)*, Miami Beach, FL.

Rao, A. S., & Georgeff, M. P. (1996). BDI agents: From theory to practice. *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS-96)*, San Francisco.

Ratsimor, O.V., Finin, T., Joshi, A., & Yesha, Y. (2003). eNcentive: A framework for intelligent marketing in mobile peer-to-peer environments. *Proceedings of the 5th International Conference on Electronic Commerce (ICEC 2003)*, Pittsburgh, PA.

Rheingold. H. (2002). *Smart mobs: The next social revolution.* Perseus.

Ricci, F., Nguyen, Q., & Cavada, D. (2004). On-tour interactive travel recommendations. *Proceedings of the 11th International Conference on Information and Communication Technologies in Travel and Tourism (ENTER 2004)*, Cairo, Egypt.

Rudström, Å., Svensson, M., Cöster, R., & Höök, K. (2004). MobiTip: Using Bluetooth as a mediator of social context. *Adjunct Proceedings of Ubicomp 2004,* Nottingham, UK.

Sarwar, B., Karypis G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on the World Wide Web (WWW '01)*, Hong Kong.

Shekar, S., Nair, P., & Helal, A. (2003). iGrocer—A ubiquitous and pervasive smart grocer shopping system. *Proceed-*

ings of the ACM Symposium on Applied Computing (SAC), Melbourne, FL.

KEY TERMS

CDMA: Code division multiple access.

GPS: Global positioning system.

J2ME: Java Platform, Micro Edition.

PDA: Personal digital assistant.

UNSPSC: United Nations Standard Products and Services Code.

XML: Extensible Markup Language.

Integrating Pedagogy, Infrastructure, and Tools for Mobile Learning

David M. Kennedy
Hong Kong University, Hong Kong

Doug Vogel
City University of Hong Kong, Hong Kong

INTRODUCTION

With the advent of the Web, students are empowered with environments that support a wide variety of interactions. These include engagement with authentic tasks, using a range of learning resources, and engaging with teachers and/or other students in knowledge-building communicative interactions. However, the concept of the fully wired world where students can learn anytime/anywhere is still unrealized. Instead, the growth of wireless networks has been substantial, with some countries limiting the construction of wired environments in preference to wireless connectivity. Thus, student learning environments and student expectations for convenience and flexibility are evolving to include wireless solutions along with wired Internet access at home or university.

A key issue associated with the growth of wireless services is the corresponding trade-off of service quality compared to wired computing (Associated Press, 2005). The availability of services is perceived as more important than high bandwidth and high security. The growth of wireless networks in the past 10 years has been spectacular, with a raft of technologies and standards arising to provide connectivity (Fenn & Linden, 2005). There is one note of caution: one of the leaders of research into wireless technologies, Cornel University in the USA, believes that due to competing technologies, even a fully wireless campus is still some time away (Vernon, 2006). This is due to:

- limitations in the interoperability of different wireless systems;
- high power requirements of the 802.11 wireless standard necessitating powerful (heavy) batteries for PDAs and smart phones;
- lower security than wired links; and
- potential interference, resulting in frustrated users.

Therefore, the concept of the fully wired or wireless connected world is still unrealized and will remain so for some time to come. Instead, the creation of local wireless hotspots has been suggested as a more cost-effective method (in the long term) for providing greater connectivity *and flexibility* to students (Boerner, 2002). Local wireless networking is already providing wireless links for students at cafés, shopping centers, airports, schools, and universities.

American students now have very high expectations that wireless will be available in all locations on a campus (Green, 2004). In Hong Kong, government statistics list the ownership of mobile devices as having reached an extraordinary level of 122.6% of market penetration (Office of the Telecommunications Authority of Hong Kong, 2005). Students may have a multitude of mobile devices, from mobile phones, iPods, digital cameras, and personal digital assistants (PDAs), to laptop computers. It would be remiss of teachers if they did not attempt to make the effort to utilize such pervasive technologies for teaching and learning as students increasingly try to "cram learning into the interstices of daily life" (Sharples, Taylor, & Vavoula, 2005, p. 58).

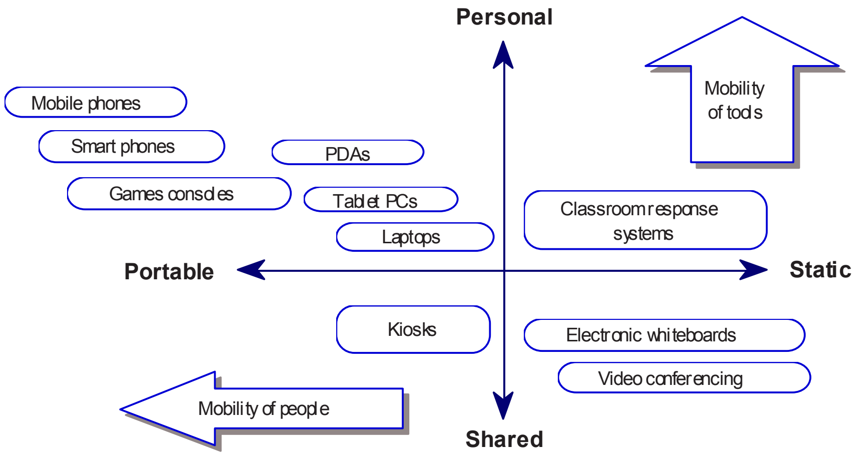
In this article the approach adopted for the design of m-learning tools and infrastructure is predicated on the idea that there will be intermittent wireless connectivity with limited bandwidth (e.g., grainy video at best). Students in Hong Kong (like most places) lead busy lives, and access to always-on Internet connections may not be possible or desired. Instead, the concept of flexibility of learning—learning at a time most suitable to the student—is seen as a primary driving factor for our work. What follows is a description of a framework for development of learning tools and institutional infrastructure designed to take advantage mobile devices and the flexibility offered by m-learning.

MOBILE DEVICES AS SEMIOTIC TOOLS

Tools were once seen as some form of some physical object (e.g., a screwdriver, the pulley, a hammer, or the cogs on a bicycle). The purpose of tools was to enhance human strength and/or human capabilities. Traditional learning included the humble pen-and-paper or an abacus. However, humans have also created semiotic tools (Vygotsky, 1978), which are intangible tools to mediate cognition. These semiotic tools include language, numbers, algebraic notation, mnemonic

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Classification of mobile technologies (Adapted from Naismith, Lonsdale, Vavoula, & Sharples, 2005, p. 7)



techniques, graphs, and diagrams—most of which may be expressed in the form of media elements that are easily stored, retrieved, and manipulated by computers (Kennedy, 2001). Since Vygotsky's time, technical tools (computers, PDAs, mobile phones, smart phones) have come to encompass devices that can utilize and manipulate signs (intangible tools) to enhance human cognitive processes (Duffy & Cunningham, 1996; Jonassen & Reeves, 1996). Current mobile devices function as computer-based cognitive tools, helping people to store, organize, structure, communicate, annotate, capture information, play, and engage in increasingly complex tasks, blurring the distinction between tangible (hardware) and intangible (software and signs) tools—one without the other is meaningless.

The feature set of mobile devices is improving rapidly as the power of the central processing unit (CPU) increases, following Moore's Law as desktop computers have for past decades (Zheng & Ni, 2006a). The future looks bright with the convergence of personal digital assistants, mobile telephones, and digital imaging into devices described as smart phones (Zheng & Ni, 2006b). The growth of computing power in such devices offers many opportunities for learning. Already such devices are endowed with features and facilities in the realm of science fiction just a few years ago, running a variety of operating systems with support for the .NET framework from Microsoft, Java, multimedia capability, and storage capacity in the multi-gigabyte realm, rapidly overcoming limitations described only a few years ago by Csete, Wong, and Vogel (2004). For example, connection speeds have risen dramatically.

However, if the potential of mobile tools for learning is not to be wasted, there is a pressing need to develop appropriate learning tools that can provide structure to the student experience. Such learning tools need to be pedagogically sound, offer high levels of interactivity, and be compliant with the available infrastructure. It has been shown that placing content on the Web or storing it in a learning management system (LMS) is not sufficient for learning to occur (Ehrmann, 1995; Reeves, 2003; Rehak & Mason, 2003). It is even more disadvantageous to do so in a mobile environment with limitations on screen size, battery life, and processing power, notwithstanding the rapid development of functionalities and features. Some examples are the virtual keyboard (http://www.virtual-laser-keyboard.com/) and more powerful batteries that enable faster, power-demanding CPUs and hard drives to be used for longer periods of time (http://www.medistechnologies.com/).

DEVELOPING FOR THE MOBILE LEARNER

Vavoula and Sharples (2002) suggested that mobility is an intrinsic property of learning. They argue that learning has spatial (workplace, university, home), temporal (days, evenings, weekends), and developmental components (the learning needs/life skills of individuals which change depending upon age, interest, or employment). Figure 1 is a diagrammatic representation of this view.

In Figure 1 there are two arrows. The horizontal arrow indicates increasing mobility of people (right to left), while the vertical arrow indicates increasing mobility of the device. In the work described in this article, the focus is on the top left quadrant, with high mobility for people and devices. Applications (mobile learning tools or m-learning applications) that support mobility of devices and people have a number of criteria that differ widely from the desktop environment. In Table 1 the basic design elements suggested by Zheng

Table 1. Design elements for user interface design for mobile learning (Adapted from Zheng & Ni, 2006a, p. 473)

Design Element	Description	Guide to UID for Mobile Learning
Context	Functional design that accounts for screen size, processor speed, educational needs	Set of icons that simplify the UI, legible font size, and adaptation of the desktop environment to suit the limitations of the mobile platform
Content	Resources that can be presented, annotated, queried, and answered	Limits on text length, image size, length of input required, use of menu-driven options for data input
Community	Information sharing (Bluetooth and WiFi)	Text, instant messaging, image messaging
Customization	Tailoring the device to the personal needs of the student	Linking to on-campus resources
Communication	Human-computer interaction	A variety of input methods, wizard-like dialogues
Connection	A variety of connection methods need to be supported	Customization of connection settings

and Ni (2006a) for user interface design (UID) of mobile devices are shown.

However, the elements shown in Table 1 neglect the all important pedagogical perspective needed to create flexible learning environments that incorporate m-learning effectively. Richards (2004) has provided a way of simplifying the paradigm wars that often plague discussions about the use of information and communication technologies (ICTs) in education by coining the expressions "new learning" and "old learning". Effectively, new learning (student-centered) is based around a social constructivist view, in which ICTs (including m-learning) are intended to:

- engage students actively in the construction of knowledge;
- involve problem solving rather than solving problems (algorithms);
- encourage and support collaboration, articulation, and discussion;
- enhance understanding rather than rote learning;
- anchor skill development and learning tasks in meaningful, authentic, and information-rich environments;
- promote motivation by interactivity;
- promote learning through cooperative, group-based activity; and
- focus on engaging activities that require higher-level skills by integrating lower-level skills (Roblyer, Edwards, & Havriluk, 2002, p. 51).

The change from old learning to new learning also has a technological equivalent. The development of a raft of mobile devices that are both personal and portable (see Figure 1) is exemplified by the relationships articulated in Table 2. In Table 2, Sharples et al. (in press) link the changes in learning methodology and current thinking to the developments in technology that support the current focus.

In summary, mobile devices support a synergy of tools, both tangible and intangible, providing ever-increasing functionality and modes by which knowledge can be constructed, annotated, stored, shared, and created.

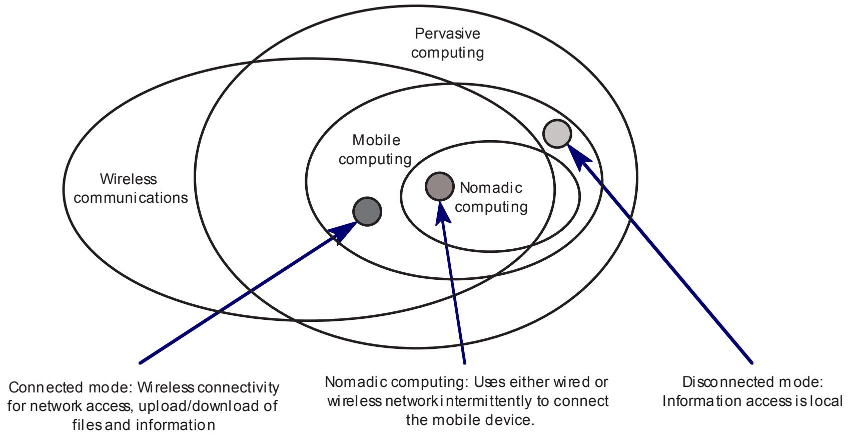
FOUR EXAMPLES OF MOBILE TOOLS

At the City University of Hong Kong, the development of mobile tools has focused on providing more flexible possibilities for learning by students. Students in Hong Kong are like most others, juggling busy lives and moving in and out of wired and wireless environments. Zheng and Ni (2006a) have suggested a model for the manner in which one may

Table 2. New learning and technology (Adapted from Sharples et al., in press)

New Learning	New Technology
Personalized	Personal
Learner-Centered	User-Centered
Situated in Time and Place	Mobile
Collaborative	Collaborative and Networked
Ubiquitous	Ubiquitous
Lifelong and Life-Wide	Connected
Activity-Based	Greater Functionality

Figure 2. Some notions of mobile computing (Adapted from Zheng & Ni, 2006a, p. 471)


Pervasive computing
Mobile computing
Wireless communications
Nomadic computing
Connected mode: Wireless connectivity for network access, upload/download of files and information
Nomadic computing: Uses either wired or wireless network intermittently to connect the mobile device.
Disconnected mode: Information access is local

consider the elements of mobile computing (see Figure 2) which is more congruent with current practice and infrastructure. The Hong Kong students move between the:

- *connected mode* (at the campus),
- *nomadic mode* (at home or connect to a desktop computer on campus or at home), and
- *disconnected mode* (on public transport, away from wired or wireless connections).

Current technology does not support *pervasive computing,* which will be reached when the hardware and the software become deeply embedded in the user's physical environment, so much so that the user may not even notice that he or she is interacting with a computing environment. The current environment, particularly in education, is still some way from this goal (Zheng & Ni, 2006a).

The current project is a holistic approach that seeks to:

1. develop and research the use of a range of tools that are designed to more readily support academic teachers in their quest to match the student learning outcomes with appropriate activities, assessment, and feedback (Kennedy, Vogel, & Xu, 2004);
2. develop the technical infrastructure that enables academics to publish activities to the Web or personal digital assistants (PDAs) or smart phones;
3. develop the technical infrastructure to allow lecturers to monitor student activity and record student learning outcomes about student interactions on their PDAs to the lecturers from within the university LMS, BlackBoard; and
4. provide advice (mainly pedagogical) and support (with examples) for developing content suitable for m-learning.

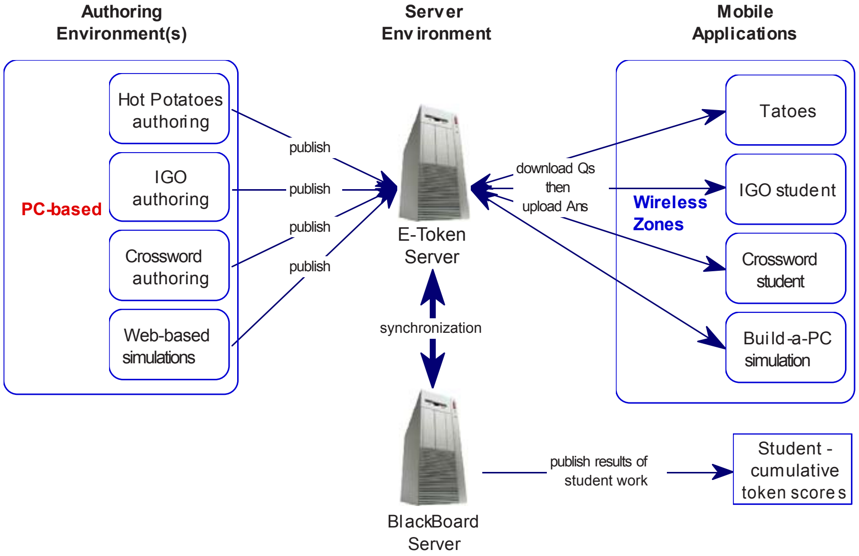
The project has been operating for nearly 18 months and is providing a degree of the flexibility demanded by students. Figure 3 provides an overview of the three components: learning tools, authoring environment, and e-token server. The concept of an e-token was suggested by tutors at City University as a method of stimulating interest and motivation among students. Students are given a number of e-tokens at the commencement of their course of study. They can use these e-tokens to buy questions. When they upload their answers and responses back to the e-token server, the log of their success or failure to solve the tasks determines how many e-tokens they receive for their efforts. An honor board of the top three students is kept on the e-token Web site.

The e-token server provides a secure connection between the university LMS (BlackBoard) and the institutional wireless network. The e-token server synchronizes with the main BlackBoard server once a day. At the time of writing, work was being undertaken to integrate this function into BlackBoard so that:

1. students can access any of the mobile resources (software and questions) directly from their BlackBoard course area (a single sign-in);
2. students upload their responses to either a personal e-portfolio or an area that is monitored by the lecturer;
3. lecturers can publish resources directly to students in either a Web-based or mobile form; and
4. lecturers can receive an instantaneous update of student downloads, uploads, and responses (rather than having to wait until the next day).

The four tools currently in use are a quiz tool based on 'Hot Potatoes' (freely available to non-profit educational institutions; http://hotpot.uvic.ca), an interactive graphing tutorial tool (IGO), a Crossword tool, and a simulation called

Figure 3. E-token server, mobile content creation, and publishing


Authoring Environment(s)
Server Environment
Mobile Applications
PC-based
Hot Potatoes authoring
IGO authoring
Crossword authoring
Web-based simulations
publish
publish
publish
publish
E-Token Server
download Qs then upload Ans
Wireless Zones
Tatoes
IGO student
Crossword student
Build-a-PC simulation
synchronization
BlackBoard Server
publish results of student work
Student - cumulative token scores

'Build a PC'. The project team has created a set of icons to provide a consistent experience for students. All of the mobile tools use the same or application-specific icons (see Figures 4, 5, and 6). The development of a set of common icons for common tasks (e.g., save, upload, download, log-in, check answer, hints, information) are consistent in each mobile application.

Tatoes

Tatoes authoring is done on a PC. Tatoes adapts quiz-based content (multiple-choice, fill-in-the-gap, ordered lists, matching exercises) created with the software Hot Potatoes to be exported to any device using the Windows .NET mobile platform. A lecturer may create a series of questions using Hot Potatoes and either publish them to the Web or in a form that students may download from the e-token server (see Figure 3). In this way instructors are only required to create one set of questions for students but have the ability to save the questions for Web or mobile delivery. One of the key pedagogical factors was to provide detailed feedback for incorrect choices in the multiple-choice questions. Charman (1999) suggests that one of the most important determinants of the success of multiple-choice questions for learning is to:

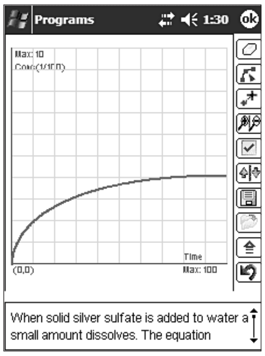
- indicate clearly if the response is correct;
- provide sufficient time for the feedback to be understood;
- consider using graphics to enhance the feedback;
- be positive, friendly, and use a neutral tone;
- provide feedback for each distractor; and
- provide references or links to other sources to improve understanding.

In the Tatoes questions, lecturers have been encouraged to adopt these principles. A sample question *Which one of the following statements best describes global IT platforms?* is shown in Table 3. Feedback regarding student use of this resource has been collected, and it will be discussed in the next section.

Table 3. An example of an MCQ item with feedback

Choice	Statement	Feedback
A:	Hardware choices are difficult in some countries because of high tariffs and import restrictions.	Yes, you are correct.
B:	Software packages are compatible in all countries when you buy from the same hardware vendor.	No, software packages developed in Europe may be incompatible with American or Asian versions. Try again.
C:	Increasing hardware costs is one of the chief reasons for the trend toward the use of global IT.	There is no evidence that hardware costs are increasing. Try again.
D:	Hardware choices are easier in global markets than in the U.S. because of short lead times for government approvals.	No, lead times for government approvals vary from country to country.

Figure 4. Mobile version of the IGO

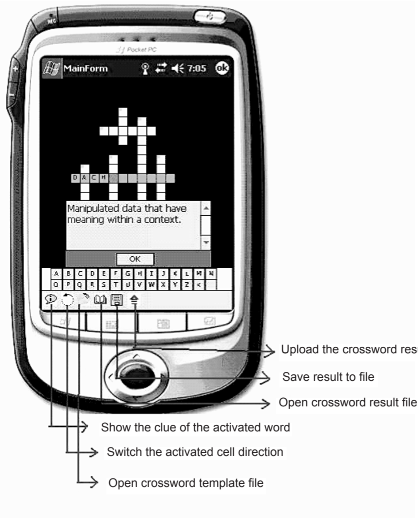


When solid silver sulfate is added to water a small amount dissolves. The equation

An initial curve has been drawn as part of a question in chemistry. The icons represent, in order:

1. Erase
2. Show handles for adjusting the shape of the curve
3. Change the format to a graph where the origin is placed centrally in the image palette (allowing graphs with positive values in the x and y axis to be displayed)
4. Shrink or grow the image
5. Next question in the sequence
6. Save a question
7. Open a question
8. Upload an answer to the e-token server
9. Go back

Figure 5. The Crossword puzzle


MainForm
7:05
ok
OK
Upload the crossword result
Save result to file
Open crossword result file
Show the clue of the activated word
Switch the activated cell direction
Open crossword template file

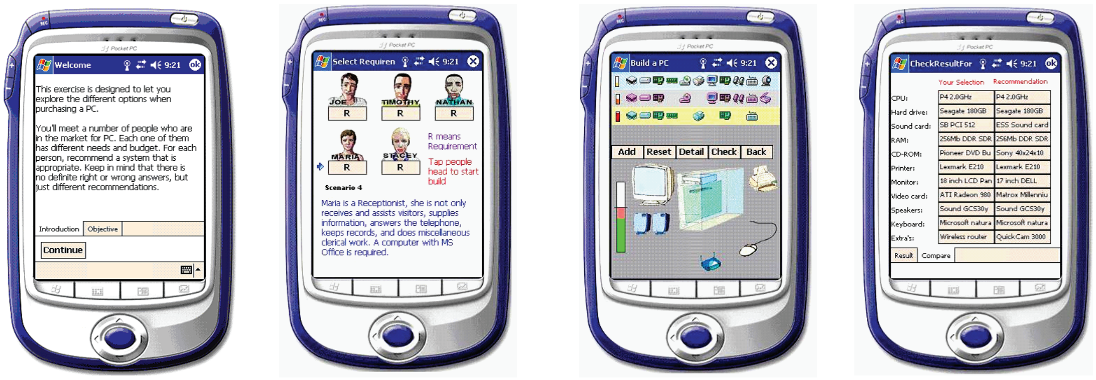
Manipulated data that have meaning within a context.

The Crossword puzzle uses the same basic icons set as all other applications being developed for this project. It has proved to be very popular with students. However, there are application-specific icons including "show a clue for the current selection" and "switch the active cell direction." The use of icons and minimal descriptions and hints has provided an interface that students do not find difficult to engage with.

Interactive Graphing Object (IGO)

Graphs are used for a large number of subjects to provide a visual representation between two or three variables. Moreover, many concepts are best represented by graphs. Achieving understanding of key concepts in science, business, and medical sciences is frequently more efficacious using graphs or other visual representations (Kremer, 1998; Kennedy, 2002; Kozma, 2000). The interactive graphing object (IGO) was originally developed for the Web (Kennedy, 2004). The IGO enabled students to sketch a graph onscreen in order to articulate their knowledge directly. This is in contrast to watching an animation or choosing the correct graph from a set of static images. The current project has created a Java-based version of the IGO that operates on either the Palm PDA or a PDA with the Windows Mobile operating system (Kennedy et al., 2004). A screen capture of the mobile version is shown in Figure 4. The IGO authoring environment enables complex questions and feedback involving graphical representations to be created for either the Web or the PDA. At this time there is only limited evaluation data available for this highly interactive tool.

Figure 6. The Build-a-PC mobile interface; (a) the introduction, (b) the role interface of Build-a-PC, (c) interactive exercise, and (d) expert comparison



This exercise is designed to let you explore the different options when purchasing a PC.

You'll meet a number of people who are in the market for PC. Each one of them has different needs and budget. For each person, recommend a system that is appropriate. Keep in mind that there is no definite right or wrong answers, but just different recommendations.

Maria is a Receptionist, she is not only receives and assists visitors, supplies information, answers the telephone, keeps records, and does miscellaneous clerical work. A computer with MS Office is required.

	Your Selection	Recommendation
CPU:	P4 2.0GHz	P4 2.0GHz
Hard drive:	Seagate 180GB	Seagate 180GB
Sound card:	SB PCI 512	ESS Sound card
RAM:	256Mb DDR SDR	256Mb DDR SDR
CD-ROM:	Pioneer DVD Bu	Sony 40x24x10
Printer:	Lexmark E210	Lexmark E210
Monitor:	18 inch LCD Pan	17 inch DELL
Video card:	ATI Radeon 980	Matrox Millenniu
Speakers:	Sound GCS30y	Sound GCS30y
Keyboard:	Microsoft natura	Microsoft natura
Extra's:	Wireless router	QuickCam 3000

(a) (b) (c) (d)

Crossword Puzzle

The Crossword tool creates crossword puzzles with hints and clues for the PDA. It is an application that has been designed to operate specifically on the PDA and does not have a Web version. Two screen captures of the interface are shown in Figure 5.

Build-a-PC Tool

The Build-a-PC has both a Web form and a mobile form (similar to the IGO). The four elements of the user interface are shown in Figures 6A to 6D. The task involves the student in an authentic task: to select the appropriate desktop system computer system for a range of people in an organization. Students must make decisions based upon the position held, responsibilities of the individual in the organization, and the budget available for the purchase.

At this stage the tool is available and has been used in two courses. However, the programming to collect log data is not yet in place.

In Figures 6(a), 6(b), and 6(c)), the interface for the activity is shown. Figure 6(a) shows the introduction to the task where a student is told that she or he is the purchasing officer for a company. Figure 6(b) provides the list of company personnel and a description of their positions in the company for whom the student must decide which configuration is most cost efficient and effective. Figure 6(c) is the interactive portion is which students identify components on the top lines and insert into the computer configuration. Feedback is available in text form relative to task completion as well as a ‘budget bar’. At the conclusion of the task, the student can compare his or her choices against the expert view (Figure 6(d)).

CURRENT EVALUATION DATA OF TATOES, IGO, AND CROSSWORD

Both qualitative and quantitative evaluations are underway with the tools that have been developed. To date, perceptual data has been gathered from two courses of study. The first is from 416 students in multiple sections of an introductory business course (of approximately 800 students) in Semester 1, 2005-2006 year, and the second an introductory business class in Semester 2 of 2006. The use of the PDA was not mandated in either course, but was a voluntary activity. Current feedback from the tutors indicates that students who engaged with the use of mobile devices for learning found the experience of using a PDA to be a significantly better learning experience. Currently 186 students have used the e-token system from a total of 812 enrolled in the Semester 2 course. The use of the e-token system is entirely voluntary. Results are encrypted to discourage inappropriate sharing of results. Evaluation of each of the tools will be discussed separately.

Tatoes

A total of 87 students downloaded Tatoes-based exercises. However, only 28 students uploaded results back to the server. This evaluation is based upon analysis of this data. Only one student (of those who uploaded data) actively explored alternative answers after arriving at the correct answer—looking at the final distracter of one question after making three attempts to arrive at the correct answer. All other students stopped exploring alternatives once they discerned the correct answer (sometimes after one, two, or even three wrong answers first). This is in contrast to a study undertaken by Fritze (1994) with undergraduate chemistry students. In that study, there was evidence that students will take the opportunity to explore explanations provided as feedback to alternatives in problems. Fritze (1994) provided a method of visually mapping problem-solving strategies from which he identified a number of problem exploration strategies, some of which involved students repeatedly examining hints and explanations. However, this strategy was not observed in this pilot study. There is some evidence that the approach adopted by students in Hong Kong depends to a significant degree on how the assessment task is perceived (Tang & Biggs, 1996). If a novel situation or task is given without sufficient guidance (e.g., a set of new strategies or a framework of undertaking the task), the students tend to revert to established patterns of coping "in a highly surface-oriented assessment environment" (Tang & Biggs, 1996, p. 179). So while the students may wish to adopt a deep approach to learning, the perception of an assessment task may prevent them from doing so.

The current data bears this out. The pedagogical approach adopted when the questions were written was intended to encourage a deep approach to learning. In particular, high-quality feedback was provided to the students for all distracters (see Table 3). It is clear from the log data that students ceased exploring alternatives once they arrived at the right answer (except for the single student indicated above). Some of the other 59 students who downloaded the Tatoes exercises may have used a more exploratory approach, but this data is not available at this time (interviews were to be scheduled in the fall semester of 2006).

In the case of the students who uploaded their responses back to the server to receive e-tokens, there may have been some confusion with respect to trying to achieve the best possible mark in light of the assessment-based culture typical of Hong Kong education. The Tatoes log record does not penalize students for looking at alternatives once the correct answer has been selected. However, students may not have realized this, or having looked at the alternatives, then decided there was little point to uploading their scores back to the e-token server. Penalties are only incurred for incorrect selections leading up to the correct answer. Therefore, encouraging active use of the distracters (after achieving the correct response) to confirm or query why these are not the correct answer is a challenge for the next iteration of the project.

The Crossword Puzzle

At the time of writing, 47 students have downloaded on average eight crosswords each. This resource has proved very popular for students revising for their examinations, especially for an activity that does not specifically carry any grades. The nature of the crossword puzzle is focused on keywords and concepts. Students may compare answers with the crossword, but may not beam the answer to each other since that feature has been disabled by the Crossword software. It is expected that students will use the crossword puzzles as points of discussion and debate, rather than as an assessment target to be met.

The IGO

The IGO has been evaluated by an international panel of educational technology experts (Jones, 2004; Jones & McNaught, 2005). These evaluations have been reported elsewhere but some elements are shown here (Kennedy, 2004). Examples of key comments made by the expert reviewers are:

This project epitomizes a very important kind of learning object. It can be the basis of an unlimited number of applications across many fields in mathematics, science and social science.

The learning object supports core learning processes that are rarely dealt with in Web-based materials.

The IGO is similar to Data Works and Language Works in that it calls on students to simultaneously deal with textual/ conceptual and graphical representations. It is significantly more sophisticated in that it allows for seven different graph forms and these graphs are based on student-supplied data. Students create a graph and receive feedback on several key aspects of their work. Equally important, they are able to redo their graph in light of this feedback. The reiterative process in which the student acts, receives feedback, and acts on that feedback, all in the context of conceptual understanding, is potentially an extraordinarily powerful one. It is also different in that it is a template that allows lecturers to create their own graphs, questions and feedback without the need for a programmer or Web developer.

However, this is one of the issues of developing new technology-based tools: gaining the cooperation of lecturers who have control of the curriculum and resources. While expert evaluators were very positive, it is often difficult to implement unfamiliar tools into a curriculum, especially if the development of content is seen by lecturers as too time consuming. An earlier iteration of the IGO based upon Shockwave is still in use at The University of Melbourne (Kemm et al., 2000), and while the current version of the IGO is a more robust and user-friendly learning object to author and deploy, there remain issues of developing questions in academic disciplines: the designers need the active participation of the academic teachers to develop content-specific questions and establish a community of practice.

Table 4. Current inhibitions and future solutions to the development of mobile tools for education (Adapted from Csete et al., 2004)

Inhibitors	**Relationship to Current M-Learning Functionality**	**Future Solution**
Small screen size	Current screen size creates overlapping text and/or graphics, especially for the feedback to questions.	Flexible film display
Non-ergonomic input method	The stylus or onscreen text recognition (e.g., the graffiti script for Palm devices) limits the speed and flexibility of input an annotation.	Voice recognition Projection keyboard Cursive hand-writing recognition
Slow CPU speed	Applications run more slowly than on a desktop computer.	New breed of architecture for faster CPU
Limited memory	The size and power requirements of more powerful CPUs limit what can be placed in mobile devices.	Expansion memory card Increase internal RAM capacity
Limited battery span	Extended use is limited by current technology.	New breeds of lithium batteries or fuel cells
Ever-changing OS	This is a current and likely future problem not resolvable in the short term with many competing systems (e.g., Symbian, Windows mobile and Palm).	Open-source OS for mobile devices (e.g., Linux has been run successfully on a mobile phone)
Infrastructure compatibility	There is a plethora of wireless standards, some of which require large amounts of power.	Standards are still being developed to bridge the mobile platform.
Connectivity bandwidth	Current wireless networks have sacrificed speed for access.	3G mobile capacity and Bluetooth v.1.2 More efficient wireless protocols than currently available

THE FUTURE

The future is bright for mobile devices for education as CPU speed, battery life, and number of applications increase (see Table 4). However, what is more important is the establishment of sound pedagogical practice based upon experience and research into how students use the tools in practice, what the specific learning needs are, and how more effective feedback, communication, and collaboration can be enhanced. The three tools, Crossword, IGO, and Build-a-PC, have been designed with the facility for lecturers to provide high-quality feedback to students. The use of the Tatoes environment provides limited evidence that writing multiple-choice questions with detailed feedback may encourage some students to not only look for the correct answers to a question, but to spend time examining what makes other distracters wrong.

What this project has also made clear is the need for more support for individual academics in the development of questions for formative evaluation for students with high-quality feedback, as well as more encouragement for the students to actively explore alternatives to broaden their own understanding. In an assessment-driven educational culture, students need to be convinced that exploring explanations and feedback will not harm their overall scores, and may even help them understand the content more deeply—thus satisfying a recognized characteristic of students in Hong Kong for deep learning *and* improved exam results (Tang & Biggs, 1996).

REFERENCES

Associated Press. (2005). *Google offers free Wi-Fi for San Francisco.* Retrieved May 31, 2006, from http://www.msnbc.msn.com/id/9551548/

Boerner, G. L. (2002). The brave new world of wireless technologies: A primer for educators. *Campus Technology,* *16*(3). Retrieved May 31, 2006, from http://www.campus-technology.com/mag.asp?month=10&year=2002

Charman, D. (1999). Issues and impacts of using computer-based assessments (CBAs) for formative assessment. In S. Brown, P. Race, & J. Bull (Eds.), *Computer-assisted assessment in higher education* (pp. 85-94). London: Kogan Page.

Csete, J., Wong, Y.-H., & Vogel, D. (2004). Mobile devices in and out of the classroom. In L. Cantoni & C. McLoughlin (Eds.), *ED-MEDIA 2004* (pp. 4729-4736). *Proceedings of the 16th World Conference on Educational Multimedia and Hypermedia & World Conference on Educational Telecommunications,* Lugano, Switzerland. Norfolk VA: Association for the Advancement of Computing in Education.

Ehrmann, S. C. (1995). New technology: Old trap. *The Educom Review, 30*(5), 41-43.

Fenn, J., & Linden, A. (2005). *Gartner's hype cycle special report for 2005.* Retrieved May 31, 2006, from http://www.gartner.com/DisplayDocument?id=484424

Fritze, P. (1994). A visual approach to the evaluation of computer-based learning materials. In K. Beattie, C. McNaught, & S. Wills (Eds.), *Interactive multimedia in university education: Designing for change in teaching and learning* (pp. 273-285). Amsterdam: Elsevier.

Green, K. (2004). *The 2003 national survey of information technology in US higher education.* Retrieved May 31, 2006, from http://www.campuscomputing.net/

Jones, J. (2004). *The Interactive Graphing Object—Compiled evaluation report.* Hong Kong: The University of Hong Kong. Retrieved May 31, 2006, from http://learnet.hku.hk/production/evaluation/IGO%20-%20Display.pdf

Jones, J., & McNaught, C. (2005). Using learning object evaluation: Challenges and lessons learned in the Hong Kong context. In G. Richards & P. Kommers (Eds.), *ED-MEDIA 2005. Proceedings of the 17th Annual World Conference on Educational Multimedia, Hypermedia & Telecommunications,* Montreal, Canada (pp. 3580-3585). Norfolk VA: Association for the Advancement of Computers in Education.

Kemm, R. E., Kavnoudias, H., Weaver, D. A., Fritze, P. A., Stone, N., & Williams, N. T. (2000). Collaborative learning: An effective and enjoyable experience! A successful computer-facilitated environment for tertiary students. In J. Bourdeau & R. Heller (Eds.), *ED-MEDIA 2000* (pp. 9-20). *Proceedings of the 12th Annual World Conference on Educational Multimedia, Hypermedia & Telecommunications,* Montreal, Canada. Norfolk, VA: Association for the Advancement of Computers in Education.

Kennedy, D.M. (2001). *The design, development and evaluation of generic interactive computer-based learning tools in higher education.* Unpublished doctoral dissertation, The University of Melbourne, Australia.

Kennedy, D. M. (2002). Visual mapping: A tool for design, development and communication in the development of IT-rich learning environments. In A. Williamson, C. Gunn, A. Young, & T. Clear (Eds.), *Winds of change in the sea of learning: Charting the course digital education* (Vol. 1, pp. 339-348). *Proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education.* Auckland, New Zealand: UNITEC Institute of Technology. Retrieved May 31, 2006, from http://www.ascilite.org.au/conferences/auckland02/proceedings/papers/150.pdf

Kennedy, D. M. (2004). Continuous refinement of reusable learning objects: The case of the Interactive Graphing Object. In L. Cantoni & C. McLoughlin (Eds.), *ED-MEDIA 2004* (pp. 1398-1404). *Proceedings of the 16th World Conference on Educational Multimedia and Hypermedia & World Conference on Educational Telecommunications,* Lugano, Switzerland. Norfolk, VA: Association for the Advancement of Computing in Education.

Kennedy, D. M., Vogel, D. R., & Xu, T. (2004). Increasing opportunities for learning: Mobile graphing. In R. Atkinson, C. McBeath, D. Jonas-Dwyer, & R. Phillips (Eds.), *ASCILITE 2004: Beyond the comfort zone* (pp. 493-502). *Proceedings of the 21st Annual Conference of the Australian Society for Computers in Learning in Tertiary Education,* Perth, Western Australia. Retrieved May 31, 2006, from http://www.ascilite.org.au/conferences/perth04/procs/kennedy.html

Kozma, R. B. (2000). The use of multiple representations and the social construction of understanding in chemistry. In M. Jacobson & R. Kozma (Eds.), *Innovations in science and mathematics education: Advanced designs for technologies of learning* (pp. 11-46). Mahwah, NJ: Lawrence Erlbaum.

Kremer, R. (1998). *Visual languages for knowledge representation.* Retrieved May 31, 2006, from http://pages.cpsc.ucalgary.ca/~kremer/papers/KAW98/visual/kremer-visual.html

Naismith, L., Lonsdale, P., Vavoula, G., & Sharples, M. (2005). *Literature review in mobile technologies and learning* (11). Bristol, UK: FutureLab. Retrieved May 31, 2006, from http://www.futurelab.org.uk/download/pdfs/research/lit_reviews/futurelab_review_11.pdf

Office of the Telecommunications Authority of Hong Kong. (2004). *Key telecommunications statistics: December 2005.* Retrieved May 31, 2006, from http://www.ofta.gov.hk/en/datastat/key_stat.html

Reeves, T.C. (2003). Storm clouds on the digital education horizon. *Journal of Computing in Higher Education, 15*(1), 3-26.

Richards, C. (2004). From old to new learning: Global dilemmas, exemplary Asian contexts, and ICT as a key to cultural change in education. *Globalisation, Societies and Education, 2*(3), 399-414.

Rehak, D. R., & Mason, R. (2003). Keeping the learning in learning objects. In A. Littlejohn (Ed.), *Reusing online resources: A sustainable approach to e-learning* (pp. 20-34). London: Kogan Page.

Roblyer, M. D., Edwards, J., & Havriluk, M. A. (2002). *Integrating educational technology into teaching.* Columbus, OH: Prentice-Hall.

Sharples, M., Taylor, J., & Vavoula, G. (2005). Towards a theory of mobile learning. Mobile technology: The future of learning in your hands. *Proceedings of the 4th World Conference on M-Learning (M-Learn2005)*, Cape Town, South Africa. Retrieved May 31, 2006, from http://www.mlearn.org.za/CD/papers/Sharples-%20Theory%20of%20Mobile.pdf

Sharples, M., Taylor, J., & Vavoula, G. (in press). A theory of learning for the mobile age. In R. Andrews & C. Haythornwaite (Eds.), *Handbook of e-learning research.* London: Sage.

Tang, C., & Biggs, J. (1996). How Hong Kong students cope with assessment. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 159-182). Hong Kong: Comparative Education Research Centre & Camberwell: Australian Council for Educational Research.

Trinder, J. J., Magill, J. V., & Roy, S. (2005). Portable assessment: Towards ubiquitous education. *International Journal of Electrical Engineering Education, 42,* 73-78. Retrieved May 31, 2006, from http://www.findarticles.com/p/articles/mi_qa3792/is_200501/ai_n13507163

Vavoula, G. (2005). *A study of mobile learning practices (D4.4).* Birmingham, UK: MOBIlearn. Retrieved May 31, 2006, from http://www.mobilearn.org/download/results/public_deliverables/MOBIlearn_D4.4_Final.pdf

Vavoula, G., & Sharples, M. (2002). kLeOS: A personal, mobile, knowledge and learning organisation system. In M. Milrad, U. Hoppe, & Kinshuk (Eds.), *Proceedings of the IEEE International Workshop on Mobile and Wireless Technologies in Education (WMTE2002)* (pp. 152-156). Vaxjo, Sweden: Institute of Electrical and Electronics Engineers (IEEE).

Vernon, R.D. (2006). *Cornell data networking: Wired vs. wireless?* Retrieved May 1, 2006, from http://www.cit.cornell.edu/oit/Arch-Init/WIRELESS.pdf

Zheng, P., & Ni, L. M. (2006a). *Smart phone & next generation mobile computing.* San Francisco: Elsevier.

Zheng, P., & Ni, L. M. (2006b). The rise of the smart phone. *IEEE Distributed Systems Online 1541-4922, 7*(3). Retrieved May 31, 2006, from http://csdl2.computer.org/comp/mags/ds/2006/03/o3003.pdf

KEY TERMS

Computer-Based Cognitive Tool: One of a set of software tools that support user cognition or thinking to construct knowledge by the manipulation of signs (e.g., graphs, language, and/or mathematics) using computers.

Intangible: Not able to be physically grasped. For example, language is an intangible tool, one that supports the mediation of cognition/thinking.

M-Learning: The use of electronic mobile devices to support learning, either wholly or partially, formal or informal. M-learning is a part of e-learning that makes use of the convenience of mobile devices to provide flexibility when and where the student learns.

New Learning: Student-centered learning based upon a social constructivist view of teaching and learning, involving authentic, meaningful activities, problem solving, cooperation, collaboration, articulation, and discussion.

Semiotic Tool: One of a set of intangible tools that enable meaning to be conveyed (e.g., language, mathematics, or computer software).

Smart Phone: A device that represents the convergence of a raft of technologies, including personal digital assistants (PDAs), global positioning satellite (GPS) systems, digital cameras, mp3 and video players, and mobile telephony.

Wireless: A raft of technologies that enable the transfer of data to and from appropriately equipped devices without the aid of wires. Wireless technologies include infrared, Bluetooth, and the wireless protocols in mobile telephony.

Intelligent Medium Access Control Protocol for WSN

Haroon Malik
Acadia University, Canada

Elhadi Shakshuki
Acadia University, Canada

Mieso Kabeto Denko
University of Guelph, Canada

INTRODUCTION

This article reports an ongoing research that proposes an approach to the expansion of sensor-MAC (S-MAC), a cluster-based contention protocol to intelligent medium access control (I-MAC) protocol. I-MAC protocol is designed especially for wireless sensor networks (WSNs). A sensor network uses battery-operated computing and sensing devices. A network of these devices are used in many applications, such as agriculture and environmental monitoring.

The S-MAC protocol is based on a unique feature: it conserves battery power by powering off nodes that are not actively transmitting or receiving packets. In doing so, nodes also turn off their radios. The manner in which nodes power themselves off does not influence any delay or throughput characteristics of the protocol. Inspired by the energy conservation mechanism of the S-MAC, we unmitigated our efforts to augment the node lifetime in a sensor network. In such a network, border nodes act as shared nodes between virtual clusters. Virtual clusters are formed on the basis of sleep and wake schedule of nodes. To prolong the lifetime of the network, nodes are allowed to intelligently switch to cluster where they experience minimum energy drain. Towards this end, we propose a multi-agent system at each node. This system includes two types of agents: stationary monitoring agent (SMA) and mobile mote agent (MMA). SMA is a static agent and has the functionality to monitor the node events. MMA is a mobile agent with the ability to roam in WSN. This article focuses on the architecture of the proposed system.

BACKGROUND

Recently, there has been development and adoption of many commercial and industry wireless communication standards. WSNs are a new class of wireless networks that have appeared in the last few years. Sensor networks consist of individual nodes that are able to interact with their environment by sensing or controlling physical parameters. They perform local computation based on the data gathered and transmit the results to their neighbors. Sensor nodes need to collaborate with each other to fulfill their tasks when a single node is incapable of doing so. The range of applications of WSN is rapidly growing. Some envisaged areas of applications include monitoring the environment, assisting healthcare professionals, military communication, and precision agriculture (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002; Stojmenovic, 2006).

WSNs differ from traditional wireless networks in several ways (Akyildiz et al., 2002). First, sensor networks consist of a number of nodes and have high network density that competes for the same channel. Second, most nodes in sensor networks are battery powered, and it is often very difficult to change batteries for those nodes. Third, nodes are often deployed in an ad-hoc fashion rather than with careful pre-planning; they must then organize themselves into a communication network. Fourth, sensor networks are prone to node, network failures and self-organization. Fifth, broadcasting is the main mode of communication in sensor networks and this may cause channel contentions. Finally, most traffic in the network is triggered by sensing events, and it can be extreme at times.

These characteristics of WSNs suggest that traditional MAC protocols are not suitable for wireless sensor networks without modifications. To this end, a number of MAC protocols have been proposed for WSNs (Frazer et al., 1999; Katayoun & Gregory, 1999; Heinzelman, Chandrakasan, & Balakrishnan, 2000).

These protocols are based on different design principles, including the number of physical channels used, the way a node is notified of an incoming message, and the degree of organization. Among these protocols, sensor MAC (S-MAC) uses in-channel signaling, slot structure, and a collective listening approach per slot to reduce idle-listening problem. We believe that the design principles of S-MAC enhance the energy performance for WSNs. This motivated us to propose

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

an intelligent medium access protocol, which is an extension to sensor-MAC (S-MAC) integrating a multi-agent approach. This approach allows the border nodes to intelligently select a cluster that will experience minimum energy drain.

Many research efforts are being made to propose MAC protocol in the area of WSN that is energy efficient with minimum collision and increase of throughput (Katayoun & Gregory, 1999; Heinzelman, Chandrakasan, & Balakrishnan, 2000). The MAC layer is considered as a sub-layer of the data link layer in the network protocol stack. MAC protocols have been extensively studied in traditional areas of wireless voice and data communications. Time division multiple access (TDMA), frequency division multiple access (FDMA), and code decision multiple access (CDMA) are MAC protocols that are widely used in modern cellular communication systems (Rappaport, 1996). The basic idea of these protocols is to avoid interference by scheduling nodes onto different sub-channels that are divided either by time, frequency, or orthogonal codes. Since these sub-channels do not interfere with each other, MAC protocols in this group are largely collision-free. These protocols are called scheduled MAC protocols. Other types of MAC protocols are based on channel contention. Rather than pre-allocating transmissions, nodes compete for a shared channel, resulting in probabilistic coordination. Collision happens during the contention procedure in such systems. Classical examples of contention-based MAC protocols include ALOHA (Leonard & Fouad, 1975) and carrier sensor multiple access (CSMA) (Norman, 1985). In pure ALOHA (Norman, 1985), a node transmits a packet when it is generated, while in slotted ALOHA a node transmits at the next available slot. Packets that collide are discarded and will retransmit again. In CSMA, a node listens to the channel before transmitting. If it detects a busy channel, it delays access and retries to transmit later. The CSMA protocol has been widely studied and extended. Today, it is the basis of several widely used standards, including IEEE 802.11 (LAN/MAN, 1999).

TDMA-based protocols are effective at avoiding collisions and have a built-in duty cycle extenuating idle listening. Contention-based protocols in contrast to TDMA simplify the activities and do not require any dedicated access point in a cluster. MACAW (Mark & Randy, 1997) is an example of contention-based protocol. It is widely used in wireless sensor networks and in ad-hoc networks, because of its simplicity and robustness to hidden terminal problems. The standardized IEEE 802.11 distributed coordination function (DCF) (LAN MAN, 1999) is also an example of the contention-based protocol and is mainly built on a MACAW protocol. Most of the research work (Wei & John, 2004) showed that energy consumption of the MAC protocol is very high when nodes are in idle mode. This is due to idle listening. PAMAS (Singh & Raghavendra, 1998) provided an improvement by avoiding the overhearing among neighboring nodes.

S-MAC (Wei & John, 2004) is a slot-based MAC protocol specifically designed for wireless sensor networks. Built on contention-based protocols like IEEE 802.11, S-MAC retains the flexibility of contention-based protocols while improving energy efficiency in multi-hop networks. S-MAC implements approaches to reduce energy consumption from all the major sources of energy as idle listening, collision, overhearing, and control overhead.

S-MAC: Highlights

The smooth operation of any wireless network depends, to a large extent, on the effectiveness of the low-level medium access control (MAC) sub-layer. MAC in WSN aims to ensure that no two nodes are interfering with each other's transmissions, and deals with the situation when they do. S-MAC contention-based MAC protocol not only addresses the transmission interfering issues, but also extends its efforts in minimizing the protocol-overhead, overhearing, and idle-listening. S-MAC principle is based on locally managed synchronizations and periodic sleep-listen schedules. Neighboring nodes form virtual clusters to set up common sleep schedules. If two neighboring nodes reside in two different virtual clusters, they should wakeup at listen periods of both clusters. Schedule exchange is accomplished by periodical SYNC packet broadcasts to immediate neighbors. The period for each node to send a SYNC packet is called the synchronization period.

Figure 1 shows an example scenario for sender-receiver communication. Collision avoidance is achieved by carrier sense (CS). Furthermore, Ready to send and clear to send (RTS/CTS) packet exchanges are used or unicast type data packets. An important feature of S-MAC is the concept of message-passing where long messages are divided into frames and sent in a burst. With this technique, one may achieve energy savings by minimizing communication overhead at the expense of unfairness in medium access. Periodic sleep may result in high latency, especially in multi-hop routing

Figure 1. Sender-receiver communication

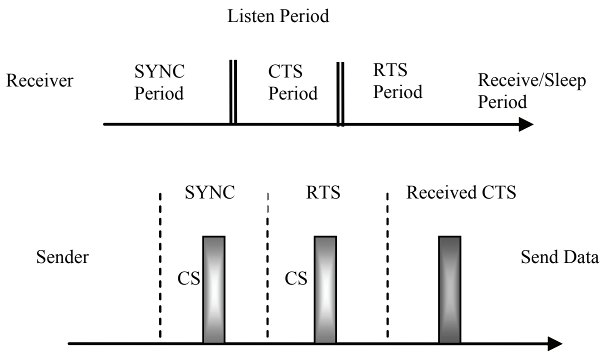
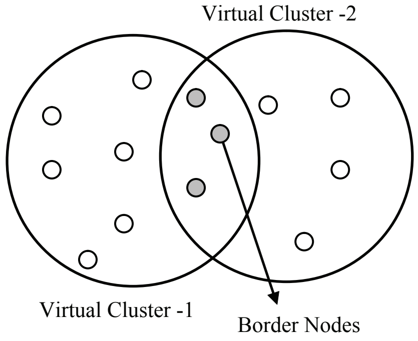


Figure 2. Border nodes



algorithm, since all immediate nodes have their own sleep schedules. The latency caused by periodic sleeping is called sleep delay (Wei & John, 2004). An adaptive listening technique is proposed to improve the sleep delay and thus the overall latency. In adaptive listening technique, the node who overhears its neighbor's transmission wakes up for a short time at the end of the transmission.

Hence, if the node is the next-hop node, its neighbor could pass data immediately. The end of the transmissions is known by the duration field of RTS/CTS packets. The energy waste caused by idle listening is reduced by sleep schedules. Broadcast data packets do not use RTS/CTS which increases collision probability. It may incur overhearing if the packets are not destined to the listening node. Sleep and listen periods are predefined and constant, which decreases the efficiency of the algorithm under variable traffic load.

One of the major problems of S-MAC protocol is the possibility of following two different schedules of neighboring nodes. This results in more energy consumptions via idle listening and overhearing. In this article, we address these problems by extending S-MAC to our proposed I-MAC protocol.

SYSTEM ARCHITECTURE

This section discusses our proposed agent-based medium access protocol for wireless sensor network, which addresses the problem of border nodes in S-MAC. This resulted when nodes try to adopt different schedules. If the radio frequency (RF) signal of a node in a cluster overlaps with the RF signal of a node in another cluster, they should have the ability to communicate with each other. If the nodes adapt the schedule of the two or more neighboring nodes, it is called border node (BN), as shown in Figure 2. Border nodes lose more energy, as they have to wake up longer than the regular nodes in a cluster. They wake up with the wake up schedule of nodes in both clusters, causing more energy drain. This raises the threat of minimizing their lifetime in the sensor network.

A node can be restricted to adopt only one schedule of its neighbors. This will stop any node from becoming border node. However, using this approach a node will remain in only one cluster for all its lifetime until it expires. In our work, a multi-agent systems approach is proposed to deal with the problem of border node, as well to provide better energy-efficient cluster. This enables the BN to join a cluster that best fits it to provide more energy efficiency. The proposed multi-agent system architecture consists of two types of agents, including stationary monitoring agent and mobile mote agent, as shown in Figure 3.

The proposed system architecture shown in Figure 3 is designed for BN to find the optimal energy-efficient cluster at given times, prolonging their lives in WSN.

The stationary monitoring agent (SMA) closely monitors the mote activity and correspondingly updates its energy model. After the border node has stationed itself to the most energy-efficient cluster, the mobile mote agent comes in to action. MMA makes use of its mobile capability by visiting and querying the energy model of BN's neighboring nodes in other virtual clusters. Thus, it periodically updates the BN about the most energy-efficient cluster it can switch.

Figure 3. System architecture

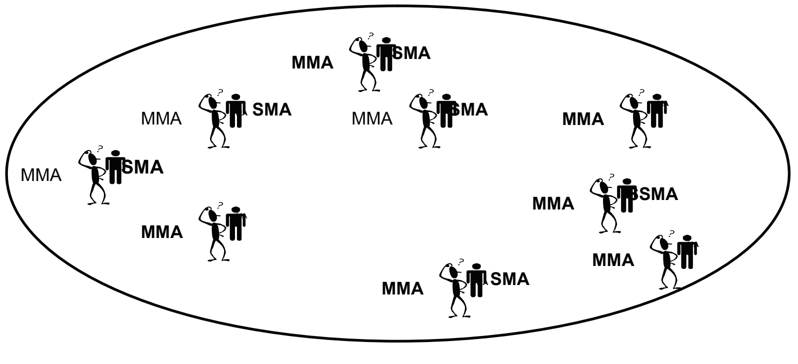


Figure 4. Agent's architecture

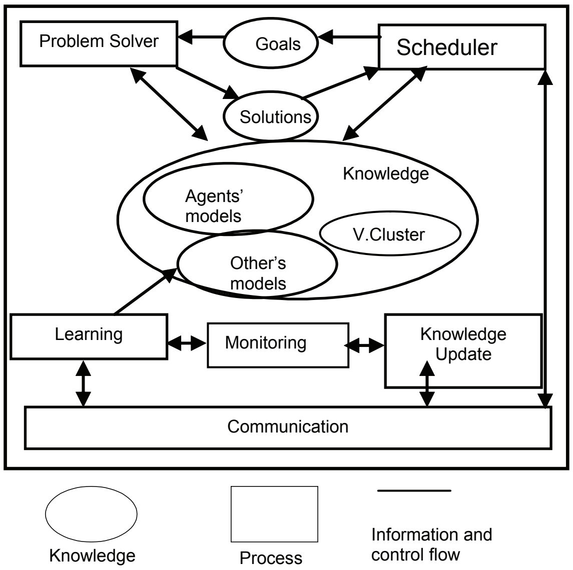
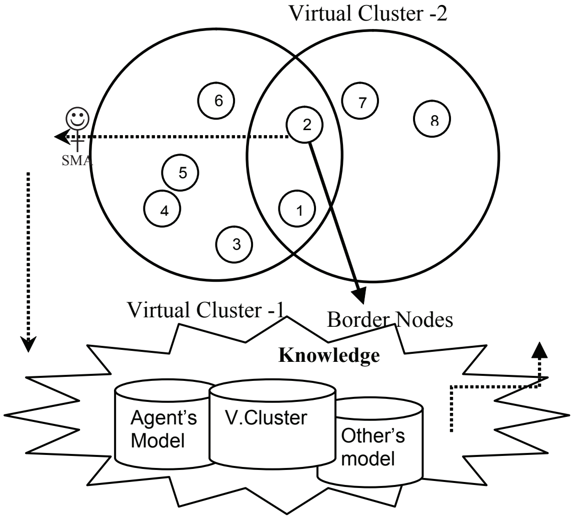


Figure 5. SMA's monitoring process



AGENT'S ARCHITECTURE

The architecture of all the proposed agents is based on the agent model proposed in Shakshuki, Ghenniwa, and Kamel (2003). Each agent poses the basic structure as shown in Figure 4, with the addition of more components based on functionality.

As shown in Figure 4, the agent architecture consists of knowledge components and executable components. The knowledge component contains the information about the WSN environment such as the number of cluster nodes it belongs to, its neighbor node, goals that need to be satisfied, and possible solutions generated to satisfy a goal. The learning component provides the agent with the capability to learn; it uses the monitored observations stored in its knowledge and runs machine-learning techniques, such as genetic algorithms, to know the energy-efficient cluster. The scheduler component provides the agent with a time agenda to start and stop certain activities such as monitoring and mobility. The communication component allows the agent to exchange messages with another agent and with an event occurring in a node. The two proposed agents (SMA and MMA) are the subject of the following two sections.

STATIONARY MONITORING AGENT

Each wireless sensor node running I-MAC protocol is equipped with a stationary monitoring agent. The SMA monitors the node and records each activity that results in loss of un-negligible energy drop. This includes transmitting data and receiving messages as data burst. The SMA closely monitors the activity of the node and records its observation into its knowledge. This includes the neighboring node ID with which the node is communicating, the virtual cluster to which the neighbor node belongs, and the amount of energy exhausted for an activity. SMA periodically updates the node's energy model.

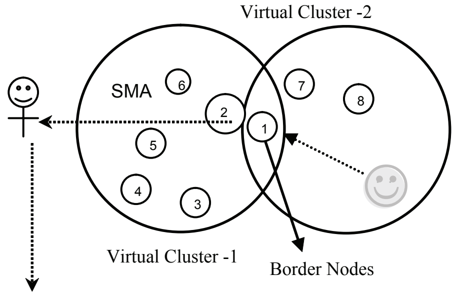
MOBILE MOTE AGENT

The mobile mote agent resides on every node in the wireless sensor network and works closely with the SMA. The MMA has the ability to travel from one node to another and benefits from knowledge acquired by its SMA that resides in the same node. In addition to moving ability, it is also able to learn using learning techniques. Use of learning techniques makes MMA scalable. As energy is a vital issue in WSN, MMA helps to learn the best energy-efficient cluster for the border node. Indeed after the border node switch to the energy-efficient cluster, the MMA still continues visiting its neighbor nodes in other virtual clusters to find the most energy-efficient cluster.

Taking advantage of the knowledge acquired by SMA shown in Figure 6, the MMA discovers virtual cluster 1 to be more energy efficient for border node 2 than virtual cluster 1. It switches the border node 2 to virtual cluster 2. Node 2 no longer remains a border node, as shown in Figure 6.

Meanwhile the SMA updates others models by marking the neighbor of node 2 in a virtual cluster as inactive. Now, node 2 will only adopt the sleep schedule of virtual cluster 2,

Figure 6. MMA on move



Id	cluster	Trans	Energy
1	1	RTS	0.00029J
2	1	CTS	0.00007J
2	1	ACK	0.0010 J

☺ MMA visiting neighbor node

☺ MMA not available at host

hence reducing its duty cycle. At this point all the monitoring events of the SMA on node 2 will belong to cluster 1.

Sensing for a particular region may increase or decrease by demand, due to the nature of the commercial application of sensor networks (Estrin, Govindan, Heidemann, & Kumar, 1999). Hence, it is vital to highlight that the node should always try to join the most energy-efficient virtual cluster at all times. Mobility of MMA now comes handy. It will visit node 2 neighbors in virtual cluster 2. It will determine the energy efficiency based on the knowledge acquired by the neighbor's SMA. In time, if the MMA finds virtual cluster 2 outperforming virtual cluster 1, it will switch node 2 to cluster 2. In doing so, the SMA will toggle the others models—that is, update by marking the neighbor in cluster 1 as inactive and in cluster 2 as active.

FUTURE TENDS

Although a number of MAC protocols have been proposed for sensor networks, several aspects still require further research. Some of the aspects and areas for further research are described as follows:

- Most existing MAC protocols for sensor networks do not support mobile sensors. Further research is needed to adapt existing MAC protocols to support mobility or to design new MAC protocols with mobility in mind.
- MAC protocols for sensor networks so far consider energy efficiency as a primary design goal. Although this is indisputably the main challenge in WSNs, future implementations and deployment need to consider performance issues such as reliable data delivery, reduced latency, and higher throughput.
- Although there are research efforts on real implementation testbeds, such work is relatively low compared to simulation-based performance evaluation of MAC protocols. In the future, performance evaluation should involve both simulation and experiments in implementation testbeds.

CONCLUSION AND FUTURE WORK

In this article, we presented a multi-agent-based system architecture to reduce the energy consumption among border nodes in S-MAC protocol. Our proposed system consists of two types of agents, including SMA and MMA. SMA monitors the events associated with a mote during communication and builds its knowledge. MMA benefits from the knowledge of SMA to predict the most energy-efficient cluster for border node at all times.

In the future, we plan to develop a simulator using Java. The environment of the simulator will include a testbed of Mica-2 motes running on Tiny OS. We also plan to deploy our agents that occupy minimal memory in motes and can easily transverse in a sensor network with fewer transmissions.

REFERENCES

Abramson, N. (1985). Development of the ALOHANET. *IEEE Transactions on Information Theory, 31,* 119-123.

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). A survey on sensor networks. *IEEE Communications Magazine, 40*(8), 102-116.

Bennett, F., Clarke, D., Evans, B. J., Hopper, A., Jones, A., & Leask, D. (1999). Piconet: Embedded mobile networking. *IEEE Personal Communications Magazine, 4*(5), 8-15.

Estrin, D., Govindan, R., Heidemann, J., & Kumar, S. (1999). Next century challenges: Scalable coordination in sensor networks. *Proceedings of ACM MobiCom'99* (pp. 263-270), Washington, DC.

Heinzelman, W. R., Chandrakasan, A., & Balakrishnan, H. (2000, January). Energy-efficient communication protocols for wireless microsensor networks. In *Proceedings of the Hawaii International Conference on Systems Sciences* (vol. 8, pp. 8020-8030), Maui, HI.

Kleinrock, L., & Tobagi, F. (1975). Packet switching in radio channels: Part I—carrier sense multiple access modes and their throughput delay characteristics. *IEEE Transactions on Communications, 23*(12), 1400-1416.

LAN/MAN. (1999). *Wireless LAN medium access control (MAC) and physical layer (PHY) specification.* Standards Committee, IEEE Computer Society.

Rappaport, T. S. (1996). *Wireless communications: Principles and practice.* Englewood Cliffs, NJ: Prentice Hall.

Shakshuki, E., Ghenniwa, H., & Kamel, M. (2003). Agent-based system architecture for dynamic and open environments. *International Journal of Information Technology and Decision Making, 2*(1), 105-133.

Singh, S., & Raghavendra, C. S. (1998). PAMAS: Power aware multi-access protocol with signalling for ad hoc networks. *ACM Computer Communication, 28*(3), 5-26.

Sohrabi, K., & Pottie, J. G. (1999). Performance of a novel self organization protocol for wireless ad hoc sensor networks. *Proceedings of the IEEE 50th Vehicular Technology Conference* (pp. 1222-1226).

Stemm, M., & Katz, K. H. (1997). Measuring and reducing energy consumption of network interfaces in hand-held devices. *IEICE Transactions on Communications, 80*(8) 1125-1131.

Stojmenovic, I. (2006). Localized network layer protocols in sensor networks based on optimizing cost over progress ratio. *IEEE Network, 20*(1), 21-27.

Woo, A., & Culler, D. (2001). A transmission control scheme for media access in sensor networks. *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking* (pp. 221-235)*,* Rome.

Ye, W., & Heidemann, J. (2004). Medium access control in wireless sensor networks. In C. S. Raghavendra, K. Sivalingam, & T. Znati (Eds.), *Wireless sensor networks* (pp. 73-92). Kluwer Academic.

KEY TERMS

Energy Model: A theoretical construct that represents the energy consumption of a sensor mote by set of variables.

Mobile Mote Agent (MMA): A mobile agent that can roam around WSN. It can query the energy model of the sensor nodes.

Stationary Monitoring Agent (SMA): A static agent designed for monitoring the node activity and energy model in the WSN.

Virtual Cluster: Refers to the logical relation existing between a set of nodes based on their wake and sleep schedules. The virtual cluster is not dependant upon geographical boundaries.

Intelligent User Preference Detection for Product Brokering

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

A good business-to-consumer environment can be developed through the creation of intelligent software agents (Maes, 1994; Nwana & Ndumu, 1996, 1997; Bailey & Bakos, 1997; Soltysiak & Crabtree, 1998) to fulfill the needs of consumers patronizing online e-commerce stores. This includes intelligent filtering services (Chanan, 2000) and product brokering services to understand users' needs before alerting users of suitable products according to their needs and preferences.

We present a generic approach to capture individual user responding towards product attributes including non-quantifiable ones. The proposed solution does not generalize or stereotype user preference, but captures the user's unique taste and recommends a list of products to the user. Under the proposed generic approach, the system is able to handle the inclusion of any unaccounted attribute that is not predefined in the system, without re-programming the system. The system is able to cater for any unaccounted attribute through a general description field found in most product databases. This is extremely useful as hundreds of new attributes of products emerge each day, making any complex analysis impossible. In addition, the system is self-adjusting in nature and can adapt to changes in a user's preference.

BACKGROUND

Although there is a tremendous increase in e-commerce activities, technology in enhancing consumers' shopping experience remains primitive. Unlike real-life department stores, there are no sales assistants to aid consumers in selecting the most appropriate product for users. Consumers are further confused by the large options and varieties of goods available. Thus there is a need to provide, in addition to the provided filtering and search services (Bierwirth, 2000), an effective piece of software in the form of a product brokering agent to understand their needs and assist them in selecting suitable products.

Definitions

A user's choice in selecting a preferred product is often influenced by the product attributes that range from price to brand name. This research shall classify attributes as accounted, unaccounted, and detected. The same attributes may also be classified as quantifiable or non-quantifiable attributes.

Accounted attributes are predefined attributes that the system is specially catered to handle. A system may be designed to capture the user's choice in terms of price and brand name, making them accounted attributes. *Unaccounted attributes* have the opposite definition, and such attributes are not predefined in the ontology of the system. The system does not understand whether an unaccounted attribute represents a model or a brand name. Such attributes merely appear in the product description field of the database. The system will attempt to detect the unaccounted attributes that affect the user's preference and consider them as *detected attributes.* Thus detected attributes are unaccounted attributes that are detected to be vital in affecting the user's preference.

Quantifiable attributes contain specific numeric values (e.g., hard disk size), and thus their values are well defined. Non-quantifiable attributes on the other hand do not have any logical numeric values, and their valuation may differ from user to user (e.g., brand name).

The proposed system shall define price and quality of a product in the ontology and consider it to be quantifiable, accounted attributes. All other attributes defined in the system and considered as unaccounted attributes will be detected by the system.

Related Work

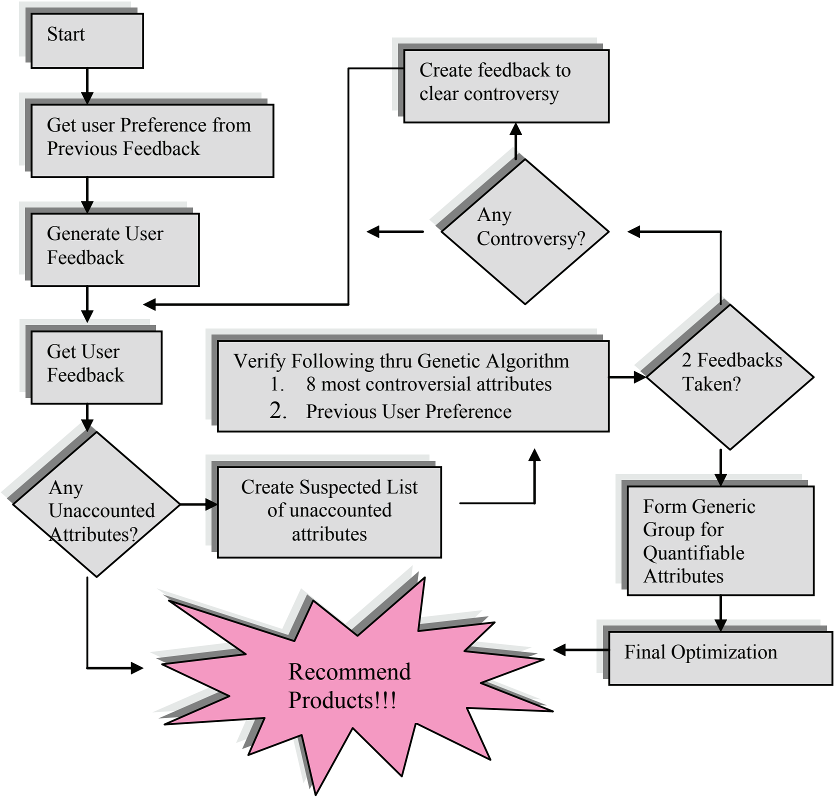
A lot of research and work has been done to aid transactions in electronic commerce. One of the research aims found is to understand a user's needs before recommending products through the use of product brokering services. Due to the difference in complexity, different approaches are proposed to handle quantifiable and non-quantifiable attributes.

One of the main approaches to handle quantifiable attributes is to compile these attributes and assign weights representing their relative importance to the user (Guan, Ngoo, & Zhu, 2002; Zhu & Guan, 2001; Sheth & Maes, 1993). The weights are adjusted to reflect the user's preference.

Much research is aimed at creating an interface to understand user preference in terms of non-quantifiable attributes. This represents a more complex problem as attributes are highly subjective with no discrete quantity to measure their

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. System flow diagram


Start
Get user Preference from Previous Feedback
Generate User Feedback
Get User Feedback
Any Unaccounted Attributes?
Create Suspected List of unaccounted attributes
Verify Following thru Genetic Algorithm
1. 8 most controversial attributes
2. Previous User Preference
2 Feedbacks Taken?
Any Controversy?
Create feedback to clear controversy
Form Generic Group for Quantifiable Attributes
Final Optimization
Recommend Products!!!

values. Different users will give different values to a particular attribute. MARI (multi-attribute resource intermediary) proposed a "word-of-mouth approach" to solve this problem. The project split up users into general groups and estimated their preference to a specific set of attributes through the group the user belongs to.

Another approach in the handling of non-quantifiable attributes involves specifically requesting the user for the preferred attributes. Shearin and Liberman (2001) provided a learning tool for the user to explore his or her preference before requesting him or her to suggest desirable attributes.

Some of the main problems in related work lie in the handling of non-quantifiable attributes, as the approaches are too general. Most work so far only attempts to understand user preference through generalization and stereotyping instead of understanding specific user needs. Another main problem is that most works are only able to handle a specific set of attributes. The attributes that they are able to handle are hard-coded into the design of the system, and the consequence is that they are not able to handle attributes that are unaccounted and beyond the pre-defined list. However, the list of product attributes is often large, possibly infinite. The approach used in related research may not be able to cover all the attributes, as they need to classify them into the ontology.

DESCRIPTION OF INTELLIGENT USER PREFERENCE DETECTION

The proposed approach attempts to capture user preference on the basis of two quantifiable accounted attributes, price and quality. It incrementally learns and detects any unaccounted attribute that affects the user's preference. If any unaccounted attribute is suspected, the system attempts to come up with a list of highly suspicious attributes and verify their importance through a genetic algorithm (Haupt & Haupt, 1998). Thus vital attributes that are unaccounted for previously will be considered. The unaccounted attributes are derived from the general description field of a product. The approach is therefore generic in nature, as the system is not restricted by the attributes it is designed to cater to.

Overall Procedure

The overall procedure is as shown in Figure 1. As the system is able to incrementally detect the attributes that affect user preference, it first retrieves any information captured regarding the user from some previous feedback and generates feedback in the form of a list of products for the user to rank, and attempts to investigate the presence of any unaccounted attribute affecting the user's preference. The system shall compile a list of possible attributes that are unaccounted for by analyzing the user feedback and rank them according to their suspicion levels. The most suspicious attributes and any information captured from previous feedback are then verified through a genetic algorithm. If two cycles of feedback are completed, the system attempts to detect any quantifiable attributes that are able to form a generic group of attributes. The system finally optimizes all information accumulated by a genetic algorithm and recommends a list of products for the user according to the preference captured.

Tangible Score

In our application, we shall consider two quantifiable attributes, price and quality, as the basis in deriving the tangible score. The effects of these two attributes are always accounted for. The equation to derive this score is as shown in equations 1-3.

Equation 1 measures the price competitiveness of the product. PrefWeight is the weight or importance the user places on price competitiveness as compared to quality with values ranging from 0 to 1.0. A value of 1.0 shows that 100% of the user's preference is based on price competitiveness. A product with a price close to the most expensive product will have a low score in terms of price competitiveness and vice versa.

Equation 2 measures the score given to quality. The quality attribute measures the quality of the product and takes a value ranging from 0.0 to 1.0. The value of "1.0 – PrefWeight" measures the importance of quality to the user.

The final score given to tangible attributes are computed by adding equations 1 and 2 as shown in equation 3 as follows:

$$\text{TangibleScore} = \text{Score}_{\text{PriceCompetitive}} + \text{Score}_{\text{Quality}} \qquad (3)$$

$$\text{Modification Score} = \left(\sum_{i=1}^{\text{NoOfAttributes}} K_i - 1\right) * \text{TangibleScore} \qquad (4)$$

Modification Score for Detected Attributes

The modification score is the score assigned to all detected attributes by the system. These detected attributes are previously unaccounted attributes, but had been detected by the system to be a vital attribute in the user's preference. These include all other attributes besides price and quality. As these attributes may not have a quantifiable value, the score is taken as a factor of the TangibleScore derived earlier. The modification score is as shown in equation 4 whereby the modification factor K is introduced.

The values of each modification factor K range between 0.0 and 2.0. A value of K shall be assigned for each newly detected attribute (e.g., each brand name will have a distinct value of K). The modification factor K takes a default value of 1.0 that gives a modification score of 0. Such a situation arises when the detected attribute does not affect the user's choice. When K<1.0, we will have a negative or penalty score for the particular attribute. This will take place when the user dislikes products from a certain brand name or other detected attributes. When K >1.0 we will have a bonus score to the attributes, and it takes place when the user has special positive preference towards certain attributes. By using a summation sign as shown in equation 4, we are considering the combined effects of all the attributes that are previously unaccounted by the system.

With all new attributes captured, the final score for the product is as shown in equation (5) as the summation of tangible and modification scores.

$$\text{Final Score} = \text{TangibleScore} + \text{Modification Score} \qquad (5)$$

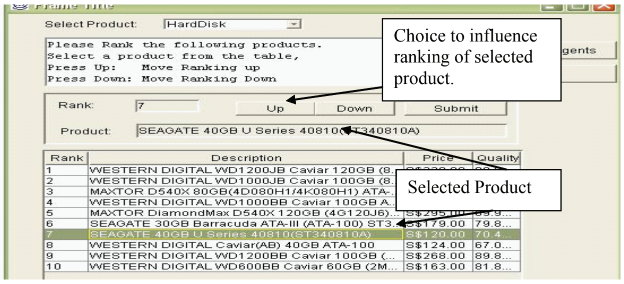
A Ranking System for User Feedback

As shown in equations 1 and 2 earlier, there is a need to capture the user's preference in terms of the PrefWeight in equation 1 and the various modification factor K in equation 4. The system shall request the user to rank a list of products as shown in Figure 2. The user is able to rank the products according to his or her preference with the up and down button and submit when done. The system shall make use of this ranked list to assess a best value for PrefWeight in equations 1 and 2. In a case whereby no unaccounted attributes affect the user's feedback, the agents will be evolved along the PrefWeight gradient to optimize a value for the PrefWeight.

Fitness of Agents

The fitness of each agent shall depend on the similarity between the agent's ranking of the product and the ranking made by the user. It reflects the fitness of agents in capturing the user's preference.

Figure 2. Requesting the user to rank a list of products



Select Product: HardDisk

```
Please Rank the following products.
Select a product from the table,
Press Up:   Move Ranking up
Press Down: Move Ranking Down
```

Rank: 7

Product: SEAGATE 40GB U Series 40810(ST340810A)

Rank	Description	Price	Quality
1	WESTERN DIGITAL WD1200JB Caviar 120GB (8...	[illegible]	[illegible]
2	WESTERN DIGITAL WD1000JB Caviar 100GB (8...	[illegible]	[illegible]
3	MAXTOR D540X 80GB(4D080H1/4K080H1) ATA-...	[illegible]	[illegible]
4	WESTERN DIGITAL WD1000BB Caviar 100GB A...	[illegible]	[illegible]
5	MAXTOR DiamondMax D540X 120GB (4G120J6)...	S$295.00	[illegible]
6	SEAGATE 30GB Barracuda ATA-III (ATA-100) ST3...	S$179.00	79.8...
7	SEAGATE 40GB U Series 40810(ST340810A)	S$120.00	70.4...
8	WESTERN DIGITAL Caviar(AB) 40GB ATA-100	S$124.00	67.0...
9	WESTERN DIGITAL WD1200BB Caviar 100GB (...	S$268.00	89.8...
10	WESTERN DIGITAL WD600BB Caviar 60GB (2M...	S$163.00	81.8...

Choice to influence ranking of selected product.

Selected Product

Unaccounted Attribute Detection

To demonstrate the system's ability to detect unaccounted attributed, the ontology shall contain only price and quality, while all other attributes are unaccounted and remain to be detected, if they are vital to the user. These unaccounted attributes include non-quantifiable attributes that are subjective in nature (e.g., brand name). The unaccounted attributes can be retrieved by analyzing the description field of a product database, thus allowing new attributes to be included without the need of change in ontology or system design.

The system firstly goes through a detection stage where it comes up with a list of attributes that affect the user's preference. These attributes are considered as unaccounted attributes as the system has not accounted for them during this stage. A *confidence score* is assigned to each attribute according to the possibility of it being the governing attribute influencing the user's preference.

The system shall request the user to rank a list of products and analyze the feedback according to the process shown in Figure 3. The agents shall attempt to explain the rankings by optimizing the PriceWeight value and various K values. The fittest agent shall give each product a score.

The system shall loop through the 10 products that are ranked by the user and compare the score given to products. If the user ranks a product higher than another, this product should have a higher score than a lower ranked product. However if the agent awards a higher score to a product ranked lower than another (e.g., product ranked 2nd has higher score than 3rd), the product is deemed to contain an unaccounted attribute causing an illogical ranking. This process shall be able to identify all products containing positive unaccounted attributes that the user has preference for.

The next step is to identify the unaccounted attributes inside these products that give rise to such illogical rankings. The products with illogical rankings are tokenized. Each word in the product description field is considered as a possible unaccounted attribute affecting the user's preference. Each of the tokens is considered as a possible attribute affecting the user's taste. The system shall next analyze the situation and modify the confidence score according to the cases as shown.

1. The token appears in other products and shows no illogical ranking: deduction of points.
2. The token appears in other products and shows illogical ranking: addition of points.

The design above only provides an estimate on the Confidence Points according the two cases described and may not be 100% reliable.

Confirmation of Attributes

Attributes captured in previous feedback may be relevant in the current feedback as the user may choose to provide more than one set of feedback. The system thus makes a hypothesis that the user's preference is influenced by attributes affecting him in previous feedback if available and eight other new attributes with the highest confidence score. The effect of these attributes on the user's preference is verified next.

Each agent in the system shall make estimation on the user's preference by randomly assigning a modification factor (or "K" value) for each of the eight attributes with high confidence score. Attributes identified to be positive are given K values greater than 1.0, while negative attributes have K values less than 1.0. The K values and PrefWeight are optimized by a genetic algorithm to improve the fitness level of the agents.

The remaining attributes undergo another filtering process whereby redundant attributes that do not affect the agent's fitness are filtered off.

Optimization Using Genetic Algorithm

The status of detected attributes perceived by the agents and the most suspicious attributes should be verified here. The PrefWeight and various K values shall be optimized here to

Figure 3. Process identifying products with illogical rankings

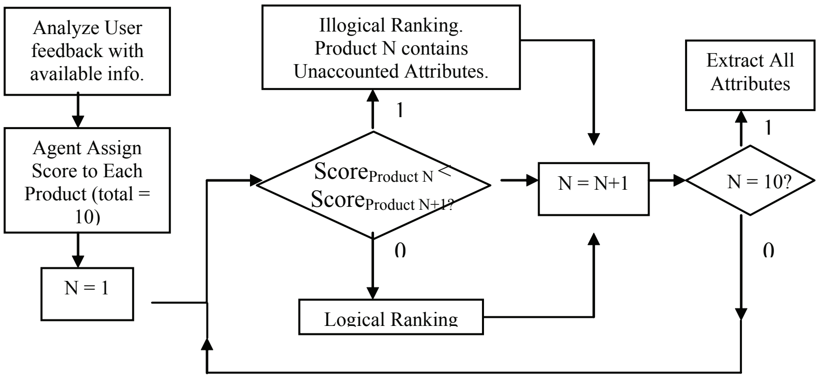

Analyze User feedback with available info.
Agent Assign Score to Each Product (total = 10)
N = 1
Illogical Ranking. Product N contains Unaccounted Attributes.
1
Score_Product N < Score_Product N+1?
0
Logical Ranking
N = N+1
N = 10?
1
0
Extract All Attributes

Figure 4. Chromosome encoding

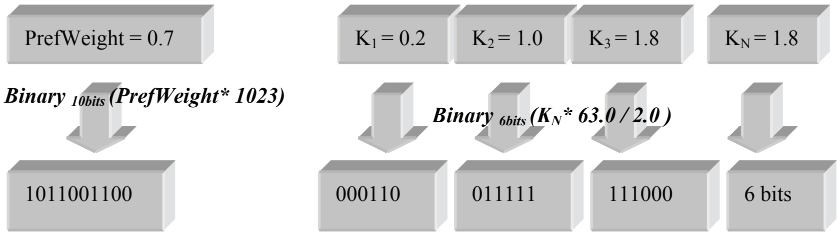
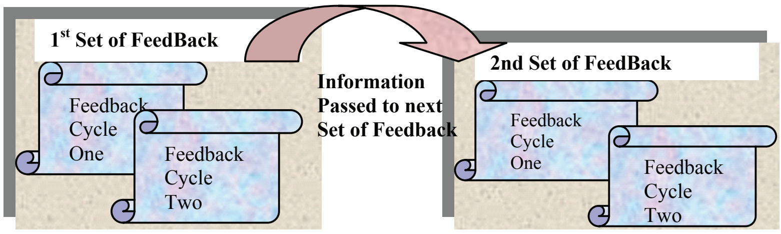


Figure 5. One set of feedback consisting of 2 feedback cycles



produce maximum agent fitness. As this represents a multidimensional problem (Osyczka, 2001) with each new K value introducing a new dimension, we shall optimize using a genetic algorithm that converts the attributes into binary strings. The agents shall evolve under the genetic algorithm to optimize the fitness of each agent. The PrefWeight and various K values are optimized in this algorithm.

The attributes of each agent are converted into a binary string as shown in Figure 4, and each binary bit represents a chromosome. In the design, 10 bits of data are used to represent the PrefWeight, while five bits of data are used to represent the various K values.

An Incremental Detection System and Overall Feedback Design

The system takes an incremental detection approach in understanding user preference, and the results show success in analyzing the complex user preference situation. The system acknowledges that not all vital attributes may be captured within one set of feedback and thus considers the results of previous sets. The attributes that affect a user's preference in one feedback become the prime candidates in the next set of feedback. In this way, the attributes that are detected are preserved and verified, while new unaccounted attributes are being detected, allowing the software agents incrementally to learn about the attributes that affect the user's preference. However some of the information captured by the system may be incorrect or no longer valid as the number of feedback cycles increase. This creates a problem in the incremental detection system, as the information may not be relevant. To solve this problem, the system checks the validity of past attributes influencing the user's preference and deletes attributes that are no longer relevant in the current feedback. Every set of feedback contains two feedback cycles as shown in Figure 6.

Both feedback cycles will attempt to detect the presence of any unaccounted attributes. In addition, the first cycle shall delete any attributes that are passed from previous feedback and no longer relevant. These attributes should have a K value of 1.0 after we apply the genetic algorithm discussed earlier. Any controversial attributes detected by the first cycle shall be clarified using the second feedback cycle.

IMPACT OF INTELLIGENT USER PREFERENCE DETECTION

A prototype was created to simulate the product broker. An independent program is written and run in the background to simulate a user. This program is used to provide feedback to the system and ranks the list of products on behalf of a simulated user who is affected by price and quality, as well as a list of unaccounted attributes. The system is also affected by some generic groups of quantifiable attributes. It was observed that the performance of the system is closely related to the complexity of the problem. More complex problems will give a lower overall performance as shown in the various cases. However, this is greatly alleviated by providing multiple sets of feedback. The system incrementally detected attributes affecting the user's preference, and in the cases shown, the gap in performance was negligible.

The system also demonstrated its ability to adapt to changes in consumer preference. This is extremely important when multiple sets of feedback are involved, as the user's preference may vary between feedback cycles. It also demonstrated the system's ability to correct its own mistakes and search for a better solution.

CONCLUSION

In this article, we demonstrated a solution in the handling of previously unaccounted attributes without the need for change in the ontology or database design. The results showed that the system designed is indeed capable of understanding the user's needs and preferences even when previously unknown or unaccounted attributes were present. The system is also able to handle the presence of multiple unaccounted attributes and classify quantifiable attributes into a generic group of unaccounted attributes.

In addition, the system demonstrated the power of incremental detection of unaccounted attributes by passing the detected attributes from within one feedback to the other.

FUTURE TRENDS

The current system generated user feedback to clarify its doubts on suspicious attributes. However, more than half of the feedback was generated in random to increase the chances of capturing new attributes. This random feedback was generated with products of different brand names having equal chances of being selected to add to the variety of the products used for feedback. This could be improved by generating feedback to test certain popular attributes to increase the detection capabilities.

REFERENCES

Bailey, J. P., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediates. *International Journal of Electronic Commerce, 1.1*(3), 7-20.

Bierwirth, C. (2000). *Adaptive search and the management of logistic systems—base models for learning agents.* Kluwer Academic.

Chanan, G., & Yadav, S. B. (2001). A conceptual model of an intelligent catalog search system. *Journal of Organizational and Electronic Commerce, 11*(1), 31-46.

Guan, S. U., Ngoo, C. S., & Zhu, F. (2002). HandyBroker—An intelligent product-brokering agent for m-commerce applications with user preference tracking. *Electronic Commerce and Research Applications, 1*(3-4), 314-330.

Haupt, R. L., & Haupt, S. E. (1998). *Practical genetic algorithm.* New York: John Wiley & Sons.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM, 37*(7), 30-40.

MIT. (n.d.). Retrieved from http://www.media.mit.edu/get-wari/MARI/

Nwana, H. S., & Ndumu, D. T. (1996). An introduction to agent technology. *BT Technology Journal, 14*(4), 55-67.

Nwana, H. S., & Ndumu, D. T. (1997). Research and development challenges for agent-based systems. *IEEE Proceedings on Software Engineering, 144*(1), 2-10.

Osyczka, A. (2001). *Evolutionary algorithms for single and multicriteria design optimization.* Physica-Verlag.

Shearin, S., & Lieberman, H. (2001). Intelligent profiling by example. In *Proceedings of the International Conference on Intelligent User Interfaces* (Vol. 1, pp. 145-151), Santa Fe, NM.

Sheth, B., & Maes, P. (1993). Evolving agents for personalized information filtering. *Proceedings of the 9th Conference on Artificial Intelligence for Applications* (pp. 345-352). IEEE Press.

Soltysiak, S., & Crabtree, B. (1998). Automatic learning of user profiles—Towards the personalization of agent services. *BT Technology Journal, 16*(3), 110-117.

Zhu, F. M., & Guan, S. U. (2001). Evolving software agents in e-commerce with GP operators and knowledge exchange. *Proceedings of the 2001 IEEE Systems, Man and Cybernetics Conference.*

KEY TERMS

E-Commerce: Consists primarily of the distributing, buying, selling, marketing, and servicing of products or services over electronic systems such as the Internet and other computer networks.

Genetic Algorithm: Search technique used in computer science to find approximate solutions to optimization and search problems. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, natural selection, and recombination (or crossover).

Ontology: Studies being or existence and their basic categories and relationships, to determine what entities and what types of entities exist.

Product Brokering: A broker is a party that mediates between a buyer and a seller.

Software Agent: An abstraction, a logical model that describes software that acts for a user or other program in a relationship of agency.

Interactive Multimedia File Sharing Using Bluetooth

Danilo Freire de Souza Santos
Federal University of Campina Grande, Brazil

José Luís do Nascimento
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

In the past few years, industry has introduced cellular phones with increasing processing capabilities and powerful wireless communication technologies. These wireless technologies provide the user with mechanisms to easily access services, enabling file sharing among devices with the same technology interfaces (Mallick, 2003). In the context of electronic commerce, which demands new techniques and technologies to attract consumers, these wireless technologies aim to simplify the shopping process and provide up-to-date information about available products.

In order to exemplify the application of mobile and wireless technologies to satisfy these new commerce functionalities and needs, we present in this article the interactive multimedia system (IMS). IMS is a system for sharing multimedia files between servers running on PCs, and client applications running on mobile devices. The system was conceived initially to be deployed in CDs/DVDs and rental stores to make available product information in a simple and interactive way.

In a general way, the system allows a user to obtain information about available products through a mobile device. Then, a user can listen or watch parts (stretches) of available videos or songs. For that, the user needs to enter the store, choose a product in the store shelf, and type its identity code in the mobile device, choosing which music (or video) to listen to (or watch).

The IMS system has a client/server architecture, where the server was developed in C++ for the Windows operating system and the client application was developed in C++ for the Symbian operating system, which is a mobile device operating system mainly used in smart phones. Client/server communication is performed based on Bluetooth wireless technology. Bluetooth is suitable for this kind of application because it has a satisfactory transmission rate with enough range, and it is also supported by more than 500 million mobile devices (Bluetooth Official Website, 2006).

The rest of this article is organized as follows. In next section we present a background of the main technologies used in this project. We then present the architecture of the proposed system and describe how the system works, before discussing future trends of mobile multimedia systems and offering final remarks.

BACKGROUND

This section provides an overview of the main technologies used in the IMS development. More specifically, we outline the Bluetooth wireless technology and the programming language used with the Symbian Operational System.

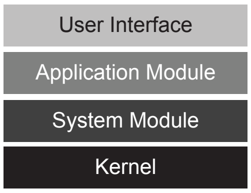
Bluetooth

To provide communication between devices, the IMS client/server architecture uses the Bluetooth wireless technology. Bluetooth is a short-range wireless technology present in a large number of smart phones of the Symbian OS Series 60 platform. It is suitable for fast file exchange, including text files, photo files, and short video files. Bluetooth technology covers a distance of about 10 meters for class 2 devices (most common devices), and each server (or master) can be connected to up to seven slaves in its coverage area (Mallick, 2003). Another important feature of Bluetooth is its lower power consumption, around 2.5 mW at most, which reinforces its use in embedded devices.

With Bluetooth, it is possible to use two kinds of connections: ACL (asynchronous connectionless) and SCO (synchronous connection oriented) (Andersson, 2001). ACL links are defined for data transmission, supporting sym-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Symbian OS architecture



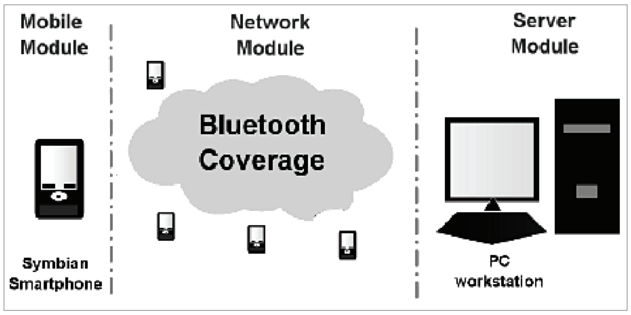
metrical and asymmetrical packet-switched connections. In this mode, the maximum data rate could be 723 kbps in one direction and 57.6 kbps in the other direction, and these rates are controlled by the master of a cell. SCO links support only a symmetrical, circuit-switched, point-to-point connection for primarily voice traffic. The data rate for SCO links is limited to 64 kbps, and the number of devices connected at the same time with the master is restricted to three devices.

In the Bluetooth protocol stack, some profiles that implement some kind of particular communication partner are defined. The general profiles in Bluetooth stack are: GAP (generic access profile), SDAP (service discovery application profile), SPP (serial port profile), and GOEXP (generic object exchange profile) (Forum Nokia, 2003). GAP defines generic procedures related to discovery of Bluetooth devices and links management aspects of connecting Bluetooth devices. SDAP defines features and procedures to allow an application in a Bluetooth device to discover services of another Bluetooth device. With SPP used in ACL links, it is possible to emulate serial cable connections using RFCOMM (RS232 Serial Cable Emulation Profile) between two peer devices. RFCOMM emulates RS-232 (Serial Cable Interface Specification) signals and can thus be used in applications that are formerly implemented with a serial cable. The GOEXP profile defines protocols and procedures that should be used by applications requiring object exchange capabilities.

Symbian

Symbian is an operating system specifically designed for mobile devices with limited resources, such as memory and processor performance. The programming language C++ for Symbian provides a specific API (application program interface), with new features for the programmer that allow access to services such as telephony and messaging (Stichbury, 2004). Also, the Symbian C++ API enables programmers to efficient deal with multitasking and memory functions. These functions reduce memory-intensive operations. Symbian OS is event driven rather than multi-thread. Although multi-thread operations are possible, they potentially create kilobytes of overhead per thread. Services in Symbian OS are provided by servers through client/server architectures. For developing applications, Symbian offers an application framework, which constitutes a set of core classes that are the basis and structure of all applications.

Figure 2. IMS general view



The Symbian OS architecture can be described by a layered approach, as illustrated in Figure 1.

The layers can be defined as follows:

- **User Interface (UI):** Can be specifically defined per vendor or per family of mobile devices, such as Series 60 platform devices;
- **Application Module:** Allows access to applications' built-in functionality concerned with data processing and not how it is presented for the user;
- **System Module:** Contains the set of OS APIs; and
- **Kernel:** The core of the operational system and cannot be directly accessed by user programs.

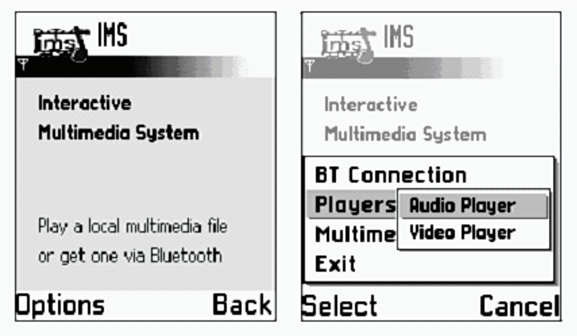
The mobile application was implemented using a Series 60 platform (Series 60 Website, 2006; Edwards & Barker, 2004). Series 60 is a complete smart phone-based UI design reference. It completes the Symbian OS architecture with a configurable graphical user interface library and a suite of applications, besides other general-purpose engines.

SYSTEM ARCHITECTURE

This section presents the IMS architecture. As introduced earlier, the IMS system has a client/server architecture, where the server was developed in C++ for the Microsoft Windows OS and the client application was developed in C++ for Symbian OS. Client/server communication is performed based on Bluetooth wireless technology, which can provide connection to up to seven users at the same time. Figure 2 illustrates a general system view.

According to this general view, the system can be divided into three specific modules: mobile, network, and server. The mobile module is composed of the software running on mobile devices, such as smart phones. It offers a friendly and intuitive user interface, which is responsible

Figure 3. IMS screenshots



Play a local multimedia file or get one via Bluetooth

for showing relevant information about available products. This information is obtained by typing the product code into the device. Then, the product description is returned with an option to run the multimedia file related to the product. Also, the mobile module controls a multimedia player installed together with the IMS software. Screenshots of the IMS mobile application are presented in Figure 3.

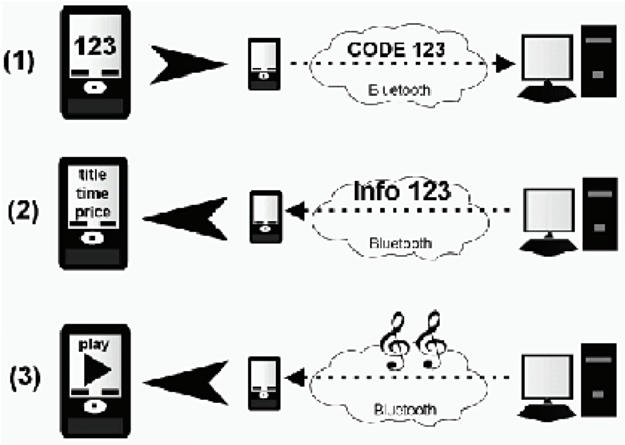
The network module controls the mobile network, in this case the Bluetooth wireless technology. To handle the connection, the network module offers an application layer protocol to manage the exchange of messages between the other two modules. This protocol is text based. Also, this module is responsible to control the transfer of files using the Bluetooth Serial Port Profile (SPP). The SPP profile is supported by Symbian (and Series 60 Platform) Bluetooth-enabled devices (Jipping, 2003). The system can support more than seven users by sharing the available Bluetooth communication ports at server side. This way, after using one communication port with one user, the network module switches the port to another user if necessary.

Finally, the server module is responsible for database and system management. Files and relevant information (description of video or song, time, author, etc.) are stored into a database. When required, the server sends relevant information about a specific product. After confirmation from the mobile software, the server sends the multimedia file.

SYSTEM EXECUTION

In this section we describe how IMS works. For this execution scenario, consider that the client has entered the IMS Bluetooth coverage area. Then, he/she has chosen a product to retrieve information (such as a DVD or CD) available at the store shelf. After that, as illustrated in Figure 4, the following steps are performed:

Figure 4. IMS operation



1. The user types the product identification code into the mobile device software. Then, the software running on the mobile device searches for the IMS server using Bluetooth technology. After finding the IMS server, the IMS network module establishes the connection between the mobile device and the server. Using a specific application protocol, the IMS software asks for information about that product code, such as audio/video stretches, release notes, and so forth.
2. If the information is available, the server module retrieves it from the database and returns it to the mobile device through the Bluetooth link. Then, the IMS software receives that information and displays it to the client, who will have available the description and the option to run a multimedia file.
3. When the client chooses to run the file, it is downloaded through the Bluetooth link and executed using a multimedia player managed by the IMS software.

FUTURE TRENDS

Today, the mobile multimedia area tends to investigate and develop mechanisms for carrying multimedia streams over wireless links, as described by Chen, Kapoor, Lee, Sanadidi, and Gerla (2004). Mobile multimedia systems are focused on providing multimedia streams directly to the mobile device in an efficient way. Future interactive multimedia systems will offer real-time multimedia streams, with more general services such as IPTV (television over IP) (Santos, Souto, Almeida, and Perkusich, 2006), radio, and so forth. Also, with the advent of smart phones with different wireless interfaces such as Wi-Fi and Bluetooth, those systems will support heterogeneous technologies offering different kinds of services for different kinds of devices.

Another prominent future trend is related to pervasive computing (Weiser, 1991; Saha & Mukherjee, 2003; Satyanarayanan, 2001). Within this context, users with mobile and wireless technology access can define their personal preferences to adapt systems and environments according to it. IMS, in the future, could be used in a pervasive infrastructure enabling file sharing according to personal preferences of users. IMS can be primarily used to delivery file in the pervasive environment, and in the future can be associated with variable bit rate multimedia streaming according to battery life of a mobile device or any kind of strategy based on the user's profile.

CONCLUSION

Mobile multimedia systems are becoming reality, and more and more accessible to ordinary users. These users, as potential consumers, are attracting industry and commerce attention, motivating research and development of several solutions in this area. Within this context, one of the relevant efforts is to promote multimedia file sharing.

To illustrate the large application of multimedia file sharing to support consumer activities, we presented in this article an interactive multimedia system. IMS provides attractive and interactive ways for users to obtain product information, which can be used in several commerce domains.

Interactive multimedia systems, such as IMS, are very useful and have great relevance for multimedia commerce. They could substitute traditional multimedia players installed in stores, offering a new way of interacting with clients using mobile devices.

REFERENCES

Andersson, C. (2001). *GPRS and 3G wireless applications.* New York: John Wiley & Sons.

Bluetooth Official Website. (2006). *Bluetooth technology benefits.* Retrieved April 28, 2006, from http://www.bluetooth.com/Bluetooth/Learn/Benefits/

Chen, L., Kapoor, R., Lee, K., Sanadidi, M.Y., & Gerla, M. (2004). Audio streaming over Bluetooth: An adaptive ARQ timeout approach. *Proceedings of the 24th International Conference on Distributed Computing Systems Workshops (ICDCSW'04).*

Edwards, L., & Barker, R. (2004). *Developing Series 60 applications—a guide for Symbian OS C++ developers.* Nokia Mobile Developer Series. Boston: Addison-Wesley.

Forum Nokia. (2003, April 4). *Bluetooth technology overview.* Retrieved March 1, 2005, from http://www.forum.nokia.com

Jipping, M. (2003). *Symbian OS communications programming.* New York: John Wiley & Sons.

Mallick, M. (2003). *Mobile and wireless design essentials.* New York: John Wiley & Sons.

Saha, D., & Mukherjee, A. (2003, March). Pervasive computing: A paradigm for the 21st century. *IEEE Computer,* 25-31.

Santos, D.F.S., Souto, S.F., Almeida, H., & Perkusich, A. (2006, April). An IPTV architecture using free software (in Portuguese). *Proceedings of the Brazilian Computer Society's Free Software Workshop in the International Free Software Forum,* Porto Alegre, Brazil.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal Communications, 8*(4).

Series 60 Web Site. (2006). *S60—about Series 60.* Retrieved April 28, 2006, from http://www.s60.com/about

Stichbury, J. (2004). *Symbian OS explained—effective C++ programming for smart phones.* New York: John Wiley & Sons.

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 94-104.

KEY TERMS

ACL: Asynchronous connectionless.

API: Application program interface.

Bluetooth Link: Connection between two Bluetooth peers.

GAP: Generic access profile.

GOEXP: Generic object exchange profile.

IMS: Interactive multimedia system.

OS: Operational system.

SCO: Synchronous connection oriented.

SDAP: Service discovery application profile.

Smart Phone: Cell phone with special computer-enabled features.

SPP: Serial port profile.

UI: User interface.

Interactive Product Catalog for M-Commerce

Sheng-Uei Guan
Brunel University, UK

Yuan Sherng Tay
National University of Singapore, Singapore

INTRODUCTION

We propose a product catalog where browsing is directed by an integrated recommender system. The recommender system is to take incremental feedback in return for browsing assistance. Product appearance in the catalog will be dynamically determined at runtime based on user preference detected by the recommender system. The design of our hybrid m-commerce catalog-recommender system investigated the typical constraints of m-commerce applications to conceptualize a suitable catalog interface. The scope was restricted to the case of having a personal digital assistant (PDA) as the mobile device. Thereafter, a preference detection technique was developed to serve as the recommender layer of the system.

BACKGROUND

M-commerce possesses two distinctive characteristics that distinguish it from traditional e-commerce: the mobile setting and the small form factor of mobile devices. Of these, the size of a mobile device will remain largely unchanged due to the tradeoff between size and portability. Small screen size and limited input capabilities pose a great challenge for developers to conceptualize user interfaces that have good usability while working within the size constraints of the device.

In response to the limited screen size of mobile devices, there has been unspoken consensus that certain tools must be made available to aid users in coping with the relatively large volume of information. Recommender systems have been proposed to narrow down choices before presenting them to the user (Feldman, 2000).

Catalog Browsing

In one study, a new user behavior, termed *opportunistic exploration,* has been identified, where users have multiple, ill-defined overlapping interests (Bryan & Gershman, 1999). Throughout the course of browsing, exposure to items affect interests, and interest may evolve due to exposure or whim. In Tateson and Bonsma (2003), the emphasis was that the paradigm of online shopping is fundamentally different from that of information retrieval.

Despite the importance of having a well-designed online catalog that supports the shopping behavior of users, the challenge of including such browsing capabilities in m-commerce is great, given that the small screen size of mobile devices severely limits the number of products that may be presented on-screen.

The predominant strategy of organizing products into narrow categories has many problems (Lee, Lee, & Wang, 2004). The alternative solution of interactive catalogs (Tateson & Bonsma, 2003) allows for fluid navigation of the product space, whereby users are given the freedom to redirect the browsing process as and when their interests change.

Recommender System

Recommender systems perform the role of sales agents by first understanding a user's preferences through querying and profiling, and subsequently presenting information or products of relevance to the user (Schafer, Konstan, & Riedl, 2001). Recommender systems have long been regarded as a highly desirable feature of e-commerce.

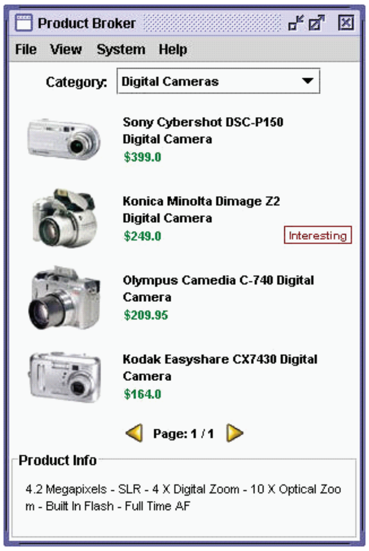
Currently there are numerous ongoing studies to improve recommender technology in the context of e-commerce (Konsten, 2004; Montaner, Lopez, & Lluis, 2003). However, the approaches of such studies are seldom directly applicable to the domain of m-commerce. With respect to the m-commerce constraints, a "best effort" recommender system that makes do with whatever information is available will serve as an interesting alternative to the "best quality" emphasis of current recommendation technology.

DESCRIPTION OF INTERACTIVE CATALOG

The interface of a catalog is divided into three components: visual presentation, browsing process, and feedback mechanism.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Screenshot



Category: Digital Cameras

Sony Cybershot DSC-P150 Digital Camera
$399.0

Konica Minolta Dimage Z2 Digital Camera
$249.0

Olympus Camedia C-740 Digital Camera
$209.95

Kodak Easyshare CX7430 Digital Camera
$164.0

4.2 Megapixels - SLR - 4 X Digital Zoom - 10 X Optical Zoom - Built In Flash - Full Time AF

Presentation

Given the constraint of a PDA screen, the main concern of our design is to maximize emphasis on product presentation while simplifying the control elements. Human cognition is more adapted to the processing of visual images as compared to textual information (Lee et al., 2004). Visual elements are thus useful mechanisms to improve the usability of a catalog. To save space while facilitating easy examination of products, we incorporate a product information panel. Figure 1 shows a screenshot of the implemented user interface.

Browsing Process

Browsing naturally induces a sense of flow, which may be imagined as a navigation process through the product space. The main challenge in the design of such a navigation system is to define the relation of products with respect to one another. Differing standpoints of people dictate that each individual sees the product relations from a different perspective. One method of custom defining product relations doing so is through interactive critiquing of products (Burke, 2002). Interactive critiquing involves allowing a user to express the goals that are not satisfied by current items. Another method to understand the preference of a user is through clustering. In our case, clustering may be used to group items that receive similar feedback from a user in an attempt to identify the underlying pattern that matches the preference of the user.

While the sharp focus on a single point in the product space, a feature of interactive critiquing, makes it unsuitable for expansive browsing, in our catalog, one desirable feature is to have an adaptable focus that allows the user to glance at the entire product range as well as zoom in on a few products of interest. We define two parameters in our browsing: breadth and preference. Breadth is a measure of diversity in the product presentation, whereas preference is the inferred interest of the user.

Breadth needs to be changed according to the state of browsing. As the user increasingly grasps some understanding of the available choices, breadth should be narrowed down to focus on recommended products based on the user's preference, allowing the user to discover products of increasing interest and at the same time facilitate a comparison of close alternatives to aid in the purchase decision. At any time, should a shift be detected in the user interest, breadth has to be relaxed accordingly to allow the user the possibility to explore again products of differing nature.

To implement such a mechanism, we divided each page of the catalog into two portions, the first containing products recommended based on the detected preference of the user and the second containing randomly sampled products. Breadth is defined as the size of the latter portion.

Feedback Mechanism

In our case, we note that the most intuitive and compact feedback method is for a user to comment directly on the products on display, as proposed by Burke, Hammond, and Young (1997), in this using a case-based critiquing approach and adopting a bipolar rating system for simplification.

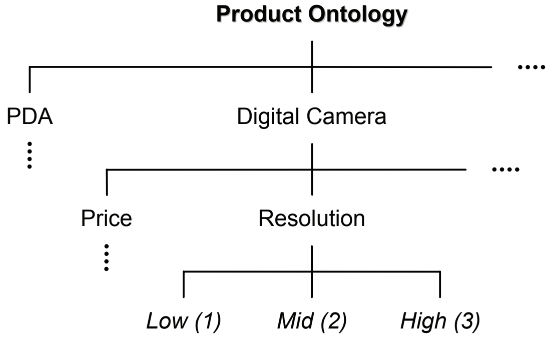
Using the bipolar rating system, we obtain a set of selected products and its complement. The selected set is derived through explicit feedback by the user. This establishes it as a strong indicator of user interest. The converse however is not necessarily true for the complementary set of non-selected products.

The usefulness of non-selected products is the relativistic nature of product selection. A user initially selects what appears to be the best available option. With greater exposure to relevant products, it is natural for a user to become more discerning in making a choice. It is thus inaccurate to conclude that non-selected products are disliked by the user. In view of the ambiguity in interpreting the set of non-selected products, the approach adopted in this article is to analyze only the selected set.

Prototype

A prototype of the catalog was developed for testing purposes. Though fully implemented in Java, the interface was designed to be easily presentable in HTML format. In an actual implementation, the catalog software is intended to reside on a Web server and be remotely accessed via PDA.

Figure 2. Product ontology



Preference Detection

Clustering is the conceptual grouping of similar products. For our case, we seek to identify a few dominant areas of interest associated with a user so to find relevant products for recommendation. To do so, we perform clustering on the set of positive examples volunteered by the user.

Product Ontology

The products are represented through the specification of an encoding scheme that maps products from the same category into a conceptual product space. The encoding scheme is responsible for the enumeration of product attributes, and in so doing, determines the relationship between products.

We adopted a static encoding scheme in the form of product ontology (Smith, 2003). In our context, product ontology is simply a descriptive tree that defines the key attributes of each product category as well as their relevant enumeration schemes. Figure 2 shows an example of a product ontology.

Product Definition

Let p denote a product and P the product space such that $p \in P$.

A product is characterized by a set of attributes as well as their associated value. We define an attribute as a particular aspect of a product's characteristics (e.g., weight, color), while an attribute instance is a value taken by a product attribute (e.g., 100g, red).

Let α denote an attribute instance and A the domain that the attribute belongs to such that $\alpha \in A$.

A product space P is defined as a vector space of η dimensions where η is the total number of unique attributes possessed by products in P.

$$\mathrm{P} : A_1 \times A_2 \times \cdots \times A_\eta$$

Products are mapped into the product space through a predefined product ontology. Products may then be represented by ordered η-tuples with the i^{th} value representing the attribute instance for the i^{th} attribute of the product. We shall refer to this η-tuple as the product characteristic.

$$p : \{\alpha_1, \alpha_2, \ldots, \alpha_\eta\},\ \alpha_i \in A_i$$

A product is assumed to be entirely characterized by the set of ordered attribute instances it is associated with.

Cluster Definition

To facilitate the clustering of products, we adopt the concept of a schema proposed by John Holland (1975) in his schema theorem. In our context, a schema is a template that partially specifies a set of product characteristics. This is possible with the introduction of wildcards that match with any value. A schema effectively defines a subset of the product space for all products that match with the schema.

Let χ denote a schema and X the schematic domain such that $\chi \in X$.

$$\mathrm{X} : \mathrm{G}_1 \times \mathrm{G}_2 \times \cdots \times \mathrm{G}_\eta \text{ where } \mathrm{G}_j = \mathrm{A}_j \cup *$$

$$\chi : \{\ \gamma_1 , \gamma_2 , \ldots , \gamma_\eta\ \} \text{ where } \gamma_j \in \mathrm{G}_j$$

To determine if a product p matches with a schema χ, we define the following functions:

$$\delta(\alpha,\gamma) = \begin{cases} 1 & \alpha = \gamma \quad or \quad \gamma = * \\ 0 & else \end{cases} \tag{1}$$

$$\delta_{match}(p,\chi) = \prod_{j=1}^{\eta} \delta(\alpha_j, \gamma_j) \tag{2}$$

A schema serves as a useful means to define a cluster, providing both a signature to determine membership to the cluster as well as a definition of product similarity. Products within a cluster are similar in the sense that they match with the schema representative of the cluster.

In this article, we shall adopt the schema as the sole definition of a cluster $\chi \equiv \mathrm{C}$. We term such an approach schematic clustering.

$$p \in C \Leftrightarrow \delta_{match}(p,\chi) = 1$$
$$p \notin C \Leftrightarrow \delta_{match}(p,\chi) = 0$$

Scoring

With the definition of cluster in place, the best cluster that generalizes a sequence of user selection S has to be found.

For this purpose, we need to be able to evaluate the relative quality of each possible cluster as a generalization of *S*.

Span

Let *S* be mapped into an $n \times \eta$ matrix $\{\alpha_{ij}\}$, such that α_{ij} denotes the j^{th} attribute instance of the i^{th} product.

Adapting the match function (2) for use on a matrix:

$$\delta_{match}(p_i, \chi) = \prod_{j=1}^{\eta} \delta(\alpha_{ij}, \gamma_j) \tag{3}$$

We define span as the number of matches a schema has on a set of products:

$$\sigma(S, \chi) = \sum_{i=1}^{n} \delta_{match}(p_i, \chi) \tag{4}$$

Given two clusters with different span, we derive greater confidence in the cluster with a larger span as an area of interest with greater significance. For example if a user selected six products, of which five belong to cluster A while only one belongs to cluster B, we naturally conclude that cluster A serves as a better representation of the user's area of interest. Span thus serves as an important measure of quality.

Order

Given a schema, we define order as the number of non-wildcard values present in the schema.

$$\overline{\delta}_{wildcard}(\gamma) = \begin{cases} 1 & \gamma \neq * \\ 0 & \gamma = * \end{cases} \tag{5}$$

$$d(\chi) = \sum_{j=1}^{\eta} \overline{\delta}_{wildcard}(\gamma_j) \tag{6}$$

Considering the definition of span, it is clear that the number of wildcards present in a schema is proportionate to the chances of the schema having a large span. However, having too many wildcards may not be a desirable because it dilutes the interpretation of the area of interest.

For example, the null schema [*,*,…,*] is undoubtedly the schema with the largest span in any situation, for it encompasses the entire product space. However, the null schema does not give any inference as to where the actual area of interest may lie. Assuming that a product fits the cluster [1,*,*,*,*] as well as the cluster [1,2,3,*,*], we see that the latter is a more precise interpretation of the area of interest because it has a more exclusive membership. Order thus serves as an equally important measure of quality as compared to span.

Span: Order Tradeoff

Having established that span and order are two competing objectives, it is not possible to maximize both measures simultaneously.

To distinguish better the quality of a schema from another, we introduce another measure called coverage:

Coverage: $$\kappa(S, \chi) = \sigma(S, \chi) \cdot d(\chi) \tag{7}$$

Score$_2$: $$\Gamma_2(S, \chi) = \kappa(S, \chi) \tag{8}$$

Coverage eliminates schemas with extreme span or order to give preference to those with a balance of the two. However, in certain cases it is still not possible to discern schemas with equally good balance. To do so, we have to decide whether to give greater priority to span or order. Since span represents a measure of the level of confidence in an area of interest, we adopt a prudent approach by giving it a higher priority.

Score$_3$: $$\Gamma_3(S, \chi) = \kappa(S, \chi) + \mu \cdot \sigma(S, \chi) \tag{9}$$

where $0 < \mu < 1$.

Noise Correction

In the context of data processing, it is usually inevitable that the data be distorted by a certain level of noise due to uncontrollable factors. In our case, noise may be introduced either due to ignorance on the part of the user or the lack of appropriate choices for the user to express freely a preference. To overcome this limitation, we introduce a noise threshold K to relax the condition for a match between a schema and a product.

$$\gamma(p_i, \chi) = \sum_{j=1}^{\eta} (1 - \delta(\alpha_{ij}, \gamma_j)) \tag{10}$$

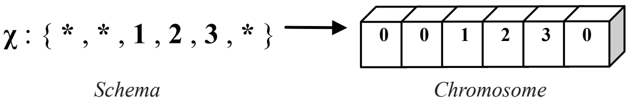
$$\delta'_{match}(p_i, \chi) = \begin{cases} 1 & \gamma(p_i, \chi) \leq K \\ 0 & \gamma(p_i, \chi) > K \end{cases} \tag{11}$$

where $0 \leq K < \eta$.

Span': $$\sigma'(S, \chi) = \sum_{i=1}^{n} \delta'_{match}(p_i, \chi) \tag{12}$$

With such an allowance given for noise, the scoring system will be able to pick up the optimum schema that matches the user preference. This is because the noise threshold allows schemas to be credited for partial matches with the selected products.

Figure 3. Genetic encoding



Owing to the noise threshold, ambiguity appears in the assessment of schemas. A schema that takes advantage of the threshold term in an unwarranted context stands to gain a higher coverage. One main reason is the simple definition of coverage as a product of span and order, which gives unnecessary credit to schema values that do not match the actual attribute instance value.

$$\delta_{pt}(\alpha,\gamma) = \begin{cases} 1 & \alpha = \gamma \quad and \quad \gamma \neq * \\ 0 & else \end{cases} \tag{13}$$

$$\textbf{Coverage:}\ \kappa'(S,\chi) = \sum_{i=1}^{n}\sum_{j=1}^{\eta}\delta'_{match}(\alpha_{ij},\gamma_j)\delta_{pt}(\alpha_{ij},\gamma_j) \tag{14}$$

With the redefinition of coverage, there is an improvement in the score to give less emphasis to matches that makes use of the noise threshold. Despite having a more equitable score, the redefined coverage is still incapable of differentiating between the sensible use of the noise threshold to accommodate noise or the abuse of it to increase coverage. To correct this error, the approach adopted is the inclusion of a penalty term to penalize the usage of the noise threshold.

$$\textbf{Penalty:}\ \pi(S,\chi) = -\sum_{i=1}^{n}\delta'_{match}(p_i,\chi)\gamma(p_i,\chi) \tag{15}$$

$$\textbf{Score:}\ \Gamma(S,\chi) = \sigma'(S,\chi) + \mu\cdot\kappa'(S,\chi) + \lambda\cdot\pi(S,\chi) \tag{16}$$

where $0 < \mu < 1$, $\lambda > 1$.

Emphasis

Finally, we recognize that a user's preference may evolve in the course of browsing. Products that were selected more recently are thus likely to be more in line with the current preference of the user. To take this factor into account, we allow a progressive emphasis to be set on more recent selection.

We define *E(i)* as the emphasis factor on a product p_i as a function of the product index in the sequence of user selection *S*. The function may follow either a linear or a geometric progression depending on the desired degree of emphasis. The emphasis factor is then applied to all application of the match function (12).

Figure 4. GA pseudocode

```
INITIALIZE population with random candidate solutions
repeat until TERMINATION CONDITION
        1.    EVALUATE chromosomes
        2.    SELECT parents
        3.    RECOMBINE pairs of parents
        4.    MUTATE offspring
        5.    EVALUATE offspring
        6.    SELECT survivors to next generation
```

The optimal emphasis varies in different contexts. Though a high degree of emphasis improves the responsiveness of the system, the tradeoff is poorer overall generalization. It is thus advisable to use moderate values of *E(i)* in most circumstances. Empirical trial tests must be carried out to investigate the effect of a chosen emphasis.

Global Optimization

Having defined a scoring function to evaluate the relative superiority of each schema, we seek to design an algorithm to search for the best schema given a sequence of user selection. EA was found to be a more appropriate choice in our context. In particular, we chose a genetic algorithm (GA), which is a form of EA for the optimization of our scoring function.

Genetic Algorithm

By assigning a value of zero to the wildcard, the η-tuple of positive integer values of a schema is encoded directly into a chromosome as an array of integers. Figure 3 illustrates the encoding process.

Having defined the chromosomes, we apply the typical genetic algorithm as summarized in Figure 4. Evaluation is done using the scoring function defined in the previous section.

Performance

To determine the performance of the algorithm, we define accuracy and efficiency as the performance measures. Accuracy is the frequency that results produce when the genetic algorithm matches the actual global optimum. We calculate accuracy as the average percentage of such matches.

On the other hand, efficiency is the amount of computational effort required to execute the algorithm. We thus calculate efficiency as the average number of generations.

CONCLUSION

The approach in this study focused on realizing the possibility for a more complete m-commerce environment. This outlook is shared by other researchers who attempt to tackle the same problem with different strategies (Guan, Ngoo, & Zhu, 2000). To the best of our knowledge, a customized catalog for m-commerce has not been conceived. This study shares the same intent to make shopping a more pleasant experience for users.

Our approach differs in the absence of a passive viewing mode, as the context of m-commerce makes it unfeasible for users to concentrate on the screen for an extended period of time. Interaction control was greatly simplified in our catalog. Through the usage of recommender technology, we streamlined the browsing process by using a reduced form of feedback.

In summary, this article highlighted the need for specialized applications in the domain of m-commerce. In particular, the need for expansive browsing as a complement to existing search and filter functions has been emphasized. As a possible solution, a novel method of product catalog navigation with the aid of a recommender system has been proposed. This approach emphasizes a minimal-attention user interface that allows a user to browse through a catalog quickly with as little cognitive effort as possible. The associated recommender system that has been conceived adopts a best effort strategy that accommodates any level of user participation. It has been shown to be capable of detecting non-linear preferences in a set of incremental feedback, as well as tolerate noisy input produced by a user. One drawback of this design is the danger of using predefined product ontology in the enumeration of attribute instances. This leads to stereotypical preference interpretation whose relevance depends largely on how the product ontology is defined.

FUTURE TRENDS

For future improvement, it may be worth investigating the possibility of having the recommender generate the ontology from the collective feedback of an ensemble of users. Another enhancement to the existing system would be to incorporate fuzzy logic into the enumeration process. Doing so eliminates the problem around segment boundaries where similar attribute values may be arbitrarily classified into different clusters.

REFERENCES

BizRate. (2004). Retrieved from www.bizrate.com

Bryan, D., & Gershman, A. (1999). Opportunistic exploration of large consumer product spaces. *Proceedings of the 1st ACM Conference on Electronic Commerce* (pp. 41-47).

Burke, R. D. (2002). Interactive critiquing for catalog navigation in e-commerce. *Artificial Intelligence Review, 18,* 245-267.

Burke, R. D., Hammond, K. J., & Young, B. C. (1997). The FindME approach to assisted browsing. *IEEE Expert: Intelligent Systems and Their Applications, 12*(4), 32-40.

Feldman, S. (2000). Mobile commerce for the masses. *IEEE Internet Computing, 4,* 75-76.

Guan, S. U., Ngoo, C. S., & Zhu, F. M. (2000). Handy broker: An intelligent product-brokering agent for m-commerce applications with user preference tracking. *Electronic Commerce Research and Applications, 1,* 314-330.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: The University of Michigan Press.

Lee, J. Y., Lee, H. S., & Wang, P. (2004). An interactive visual interface for online product catalogs. *Electronic Commerce Research, 4,* 335-358.

Montaner, M., Lopez, B., & Lluis, J. (2003). A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review, 19,* 285-330.

Schafer, J. B., Konstan, J., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery, 5,* 115-153.

Tateson, R., & Bonsma, E. (2003). ShoppingGarden—Improving the customer experience with online catalogs. *BT Technology Journal, 21*(4), 84-91.

KEY TERMS

E-Commerce: Consists primarily of the distributing, buying, selling, marketing, and servicing of products or services over electronic systems such as the Internet and other computer networks.

Feedback Mechanism: Process whereby some proportion or, in general, function of the output signal of a system is passed (fed back) to the input.

Genetic Algorithm: Search technique used in computer science to find approximate solutions to optimization and search problems. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, natural selection, and recombination (or crossover).

Global Optimization: A branch of applied mathematics and numerical analysis that deals with the optimization of a function or a set of functions to some criteria.

M-Commerce: Electronic commerce made through mobile devices.

Product Catalog: Organized, detailed, descriptive list of products arranged systematically.

Product Ontology: Studies' being or existence, and their basic categories and relationships, to determine what entities and what types of entities exist.

An Interactive Wireless Morse Code Learning System

Cheng-Huei Yang
National Kaohsiung Marine University, Taiwan

Li-Yeh Chuang
I-Shou University, Taiwan

Cheng-Hong Yang
National Kaohsiung University of Applied Sciences, Taiwan

Jun-Yang Chang
National Kaohsiung University of Applied Sciences, Taiwan

INTRODUCTION

Morse code has been shown to be a valuable tool in assistive technology, augmentative and alternative communication, and rehabilitation for some people with various conditions, such as spinal cord injuries, non-vocal quadriplegics, and visual or hearing impairments. In this article, a mobile phone human-interface system using Morse code input device is designed and implemented for the person with disabilities to send/receive SMS (simple message service) messages or make/respond to a phone call. The proposed system is divided into three parts: input module, control module, and display module. The data format of the signal transmission between the proposed system and the communication devices is the PDU (protocol description unit) mode. Experimental results revealed that three participants with disabilities were able to operate the mobile phone through this human interface after four weeks' practice.

BACKGROUND

A current trend in high technology production is to develop adaptive tools for persons with disabilities to assist them with self-learning and personal development, and lead more independent lives. Among the various technological adaptive tools available, many are based on the adaptation of computer hardware and software. The areas of application for computers and these tools include training, teaching, learning, rehabilitation, communication, and adaptive design (Enders, 1990; McCormick, 1994; Bower et al., 1998; King, 1999).

Many adapted and alternative input methods now have been developed to allow users with physical disabilities to use a computer. These include modified direct selections (via mouth stick, head stick, splinted hand, etc.), scanning methods (row-column, linear, circular) and other ways of controlling a sequentially stepping selection cursor in an organized information matrix via a single switch (Anson, 1997). However, they were not designed for mobile phone devices. Computer input systems, which use Morse code via special software programs, hardware devices, and switches, are invaluable assets in assistive technology (AT), augmentative-alternative communication (AAC), rehabilitation, and education (Caves, 2000; Leonard et al., 1995; Shannon et al., 1981; Thomas, 1981; French et al., 1986; Russel & Rego, 1998; Wyler & Ray, 1994). To date, more than 30 manufactures/developers of Morse code input hardware or software for use in AAC and AT have been identified (Anson, 1997; http://www.uwec.edu/Academic/Outreach/Mores2000/morse2000.html; Yang, 2000; Yang, 2001; Yang et al., 2002; Yang et al., 2003a; Yang et al., 2003b). In this article, we adopt Morse code to be the communication method and present a human interface for persons with physical disabilities.

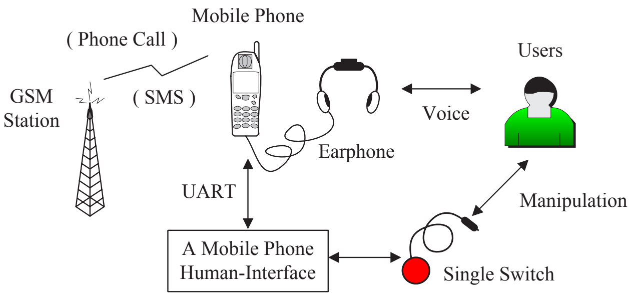
The technology employed in assistive devices has often lagged behind mainstream products. This is partly because the shelf life of an assistive device is considerably longer then mainstream products such as mobile phones. In this study, we designed and implemented an easily operated mobile phone human interface device by using Morse code as a communication adaptive device for users with physical disabilities. Experimental results showed that three participants with disabilities were able to operate the mobile phone through this human interface after four weeks' practice.

SYSTEM DESIGN

Morse code is a simple, fast, and low-cost communication method composed of a series of dots, dashes, and intervals

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. System schematics of the mobile phone human-interface



in which each character entered can be translated into a predefined sequence of dots and dashes (the elements of Morse code). A dot is represented as a period ".", while a dash is represented as a hyphen, or minus sign, "-". Each element, dot or dash, is transmitted by sending a signal for a standard length of time. According to the definition of Morse code, the tone ratio for dot to dash must be 1:3. That means that if the duration of a dot is taken to be one unit, then that of a dash must be three units. In addition, the silent ratio for dot-dash space to character-space also has to be 1:3. In other words, the space between the elements of one character is one unit while the space between characters is three units (Yang et al., 2002).

In this article, the mobile phone human interface system using Morse code input device is schematically shown in Figure 1. When a user presses the Morse code input device, the signal is transmitted to the key scan circuit, which translates the incoming analog data into digital data. The digital data are then sent into the microprocessor, an 8051 single chip, for further processing. In this study, an ATMEL series 89C51 single chip has been adopted to handle the communication between the press-button processing and the communication devices. Even though the I/O memory capacity of the chip is small compared to a typical PC, it is sufficient to control the device. The 89C51 chip's internal serial communication function is used for data transmission and reception (Mackenzie, 1998). To achieve the data communication at both ends, the two pins, TxD and RxD, are connected to the TxD and RxD pins of a RS-232 connector. Then the two pins are connected to the RxD and TxD of an UART (Universal Asynchronous Receiver Transmitter) controller on the mobile phone device. Then, persons with physical disabilities can use this proposed communication aid system to connect their mobile communication equipment, such as mobile phones or GSM (global system for mobile communications) modems, and receive or send their messages (SMS, simple message service). If they wear an earphone, they might be able to dial or answer the phone. SMS is a protocol (GSM 03.40 and GSM 03.38), which was established by the ETSI (the European Telecommunications Standards Institute) organization. The transmission model is divided into two models: text and PDU (protocol description unit). In this system, we use the PDU model to transmit and receive SMS information through the AT command of the application program (Pettersson, 2000). Structurally the mobile phone human-interface system is divided into three modules: the input module, the control module, and the display module. The interface framework is graphically shown in Figure 2. A detailed explanation is given below.

INPUT MODULE

A user's input will be digitized first, and then the converted results will be sent to the micro controller. From the signal processing circuit can monitor all input from the input device, the Morse code. The results will be entered into the input data stream. When the user presses the input key, the micro-operating system detects new input data in the data stream, and then sends the corresponding characters to the display module. Some commands and/or keys, such as *OK, Cancel, Answer, Response, Send, Receive, Menu, Exit,* and so forth, have been customized and perform several new functions in order to accommodate the Morse code system. These key modifications facilitate the human interface use for a person with disabilities.

Figure 2. Interface framework of mobile phone for persons with physical disablities

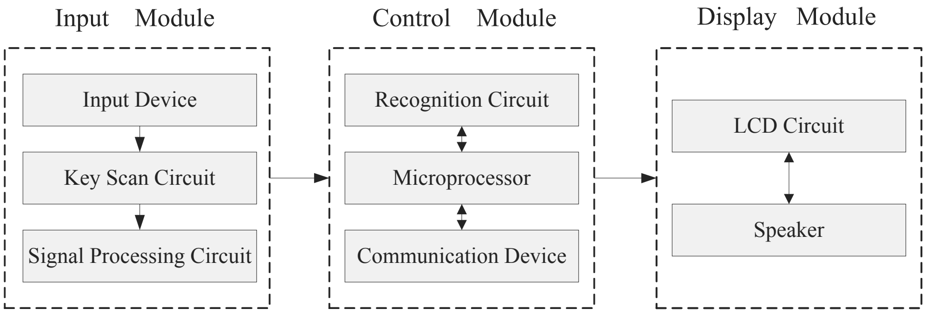
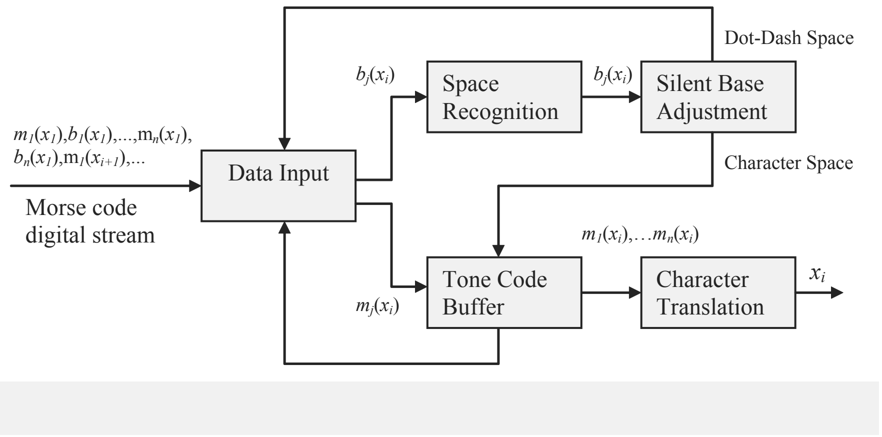


Figure 3. Block diagram of the Morse code recognition system



CONTROL MODULE

The proposed recognition method is divided into three modules (see Figure 3): space recognition, adjustment processing, and character translation. Initially, the input data stream is sent individually to separate tone code buffer and space recognition processes, which are based on key-press (Morse code element) or key-release (space element). In the space recognition module, the space element value is recognized as a dot-dash space or a character space. The dot-dash space and character space represent the spaces existing between individual characters and within isolated elements of a character respectively. If a character space is identified, then the value(s) in the code buffer is (are) sent to character translation. To account for varying release speeds, the space element value has to be adjusted. The silent element value is sent into the silent base adjustment process. Afterwards, the character is identified in the character translation process.

A Morse code character, x_i, is represented as follows:

$$m_1(x_i),\ b_1(x_i),\ \ldots,\ m_j(x_i),\ b_j(x_i),\ \ldots,\ m_n(x_i),\ b_n(x_i)$$

where

$b_j(x_i)$: jth silent duration in the character x_i.

n: the total number of Morse code elements in the character x_i.

$m_j(x_i)$: the jth Morse code element of the input character x_i.

DISPLAY MODULE

Since users with disabilities have, in order to increase the convenience of user operations, more requirements for system interfaces than a normal person, the developed system shows selected items and system condition information on an electronic circuit platform, which is based on LCD (liquid crystal display). The characteristics of the proposed system can be summarized as follows: (1) easy operation for users with physical disabilities with Morse code input system, (2) multiple operations due to the selection of different modes, (3) highly tolerant capability from adaptive algorithm recognition, and (4) system extension for customized functions.

RESULTS AND DISCUSSIONS

This system provides two easily operated modes, the phone panel and LCD panel control mode, which allow a user with disabilities easy manipulation. The following shows how the proposed system sends/receives simple message service (SMS) message or make /respond to a phone call.

SMS Receiving Operation

First, when users receive a message notification and want to look at the content, this system will provide “phone panel” and “LCD panel” control modes to choose from. In the phone panel mode, users can directly key-in Morse code “…” (as character ‘S’). The interface system will go through the message recognition process, then exchange the message into AT command “AT+CKPD=’S’, 1”, to execute the “confirm” action of the mobile phone. The purpose of this process is the same as users keying-in “yes” on the mobile phone keyboard, then keying-in Morse code “. - - . .” (as key ‘↓’). The system will recognize the message, then automatically send the “AT+CKPD=’↓’, 1” instruction. The message cursor of the mobile phone is moved to the next line, or key-in Morse code “. - . . . -” (as key ‘↑’) for moving it to the previous data line. Finally, if users want to exit and return to the previous screen, they only need to key-in Morse code ". . - ." (as character 'F'), and start the c key function on the mobile phone keyboard. If LCD panel mode is selected, one can directly follow the selected items on the LCD crystal, to execute the reception and message reading process.

SMS Transmitting Operation

Message transmission services are provided in two modes: phone panel and LCD panel. In the phone panel mode, continually type two times the Morse code ". - . - ." (as key '→'). The system will be converted into AT Command and transferred into mobile phone to show the selection screen of the message functions. Then continuing to key-in three times the Morse code ". . ." (as character 'S'), one can get into the editing screen of message content, and wait for users to input the message text data and receiver's phone number. The phone book function can be used to directly save the receiver's phone number. After the input, press the "yes" key to confirm that the message sending process has been completed. In addition, if the LCD panel mode is selected, one can follow the LCD selection prompt input the service selection of all the action integrated in the LCD panel. Then go through the interface and translate to a series of AT command orders, and batch transfer these into the mobile phone to achieve the control purpose.

The selection command "Answer a phone," displays on the menu of the LCD screen, and can be constructed using Morse code. The participants could press and release the switch, and input the number code ". - - - -" (as character '1') or hot key ". -" (as character 'A'). The mobile phone is then answered automatically. Problems with this training, according to participants, are that the end result is limited typing speed and users must remember all the Morse code set of commands.

Three test participants were chosen to investigate the efficiency of the proposed system after practicing on this system for four weeks. Participant 1 (P1) was a 14-year-old male adolescent who has been diagnosed with cerebral palsy. Participant 2 (P2) was a 14-year-old female adolescent with cerebral palsy, athetoid type, who experiences involuntary movements of all her limbs. Participant 3 (P3) was a 40-year-old male adult, with a spinal cord injury and incomplete quadriparalysis due to an accident. These three test participants with physical impairments were able to make/respond to phone calls or send/receive SMS messages after practice with the proposed system.

FUTURE TRENDS

In the future, Morse code input device could be adapted to several environmental control devices, which would facilitate the use of everyday appliances for people with physical disabilities considerably.

CONCLUSION

To help some persons with disabilities such as amyotrophic lateral sclerosis, multiple sclerosis, muscular dystrophy, and other conditions that worsen with time and cause the user's abilities to write, type, and speak to be progressively lost, it requires an assistive tool for purposes of augmentative and alternative communication in their daily lives. This article presents a human interface for mobile phone devices using Morse code as an adapted access communication tool. This system provides phone panel and LCD panel control modes to help users with a disability with operation. Experimental results revealed that three physically impaired users were able to make/respond to phone calls or send/receive SMS messages after only four weeks' practice with the proposed system.

ACKNOWLEDGMENTS

This research was supported by the National Science Council, R.O.C., under grant NSC 91-2213-E-151-016.

REFERENCES

Anson, D. (1997). *Alternative computer access: A guide to selection*. Philadelphia, PA: F. A. Davis.

Bower, R. et al. (Eds.) (1998). *The Trace resource book: Assistive technology for communication, control, and computer access*. Madison, WI: Trace Research & Development Center, Universities of Wisconsin-Madison, Waisman Center.

Caves, K. (2000). *Morse code on a computer—really?* Keynote presentation at the First Morse 2000 World Conference, Minneapolis, MN.

Enders, A., & Hall, M. (Ed.) (1990). *Assistive technology sourcebook*. Arlington, VA: RESNA Press,.

French, J. J., Silverstein, F., & Siebens, A. A. (1986). An inexpensive computer based Morse code system. In *Proceedings of the RESNA 9th Annual Conference, Minneapolis* (pp. 259-261). Retrieved from http://www.uwec.edu/Academic/Outreach/Mores2000/morse2000.html.

King, T. W. (1999). *Modern Morse code in rehabilitation and education*. MA: Allyn and Bacon.

Lars Pettersson. (n.d.). *Dreamfabric*. Retrieved from http://www.dreamfabric.com/sms

Leonard, S., Romanowski, J., & Carroll, C. (1995). Morse code as a writing method for school students. *Morsels, University of Wisconsin-Eau Claire*, *1*(2), 1.

Mackenzie, I. S. (1998). *The 89C51 Microcontroller* (3rd ed.). Prentice Hall.

McCormick, J. A. (1994). Computers and the Americans with disabilities act: A manager's guide. Blue Ridge Summit, PA: Wincrest/McGraw Hill.

Russel, M., & Rego, R. (1998). A Morse code communication device for the deaf-blind individual. In *Proceedings of the ICAART, Montreal* (pp. 52-53).

Shannon, D. A., Staewen, W. S., Miller, J. T., & Cohen, B. S. (1981). Morse code controlled computer aid for the nonvocal quadriplegic. *Medical Instrumentation*, *15*(5), 341-343.

Thomas, A. (1981). Communication devices for the non-vocal disabled. *Computer*, *14*, 25-30.

Wyler, A. R., & Ray, M. W. (1994). Aphasia for Morse code. *Brain and Language*, *27*(2), 195-198.

Yang, C.-H. (2000), Adaptive Morse code communication system for severely disabled individuals. *Medical Engineering & Physics*, *22*(1), 59-66.

Yang, C.-H. (2001). Morse code recognition using learning vector quantization for persons with physical disabilities. *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences*, *E84-A*(1), 356-362.

Yang, C.-H., Chuang, L.-Y. Yang, C.-H., & Luo, C.-H. (2002). An Internet access device for physically impaired users of Chanjei Morse code. *Journal of Chinese Institute of Engineers*, *25*(3), 363-369.

Yang, C.-H. (2003a). An interactive Morse code emulation management system. *Computer & Mathematics with Applications*, *46*, 479-492.

Yang, C.-H., Chuang, L.-Y., Yang, C.-H., & Luo, C.-H. (2003b, December). Morse code application for wireless environmental control system for severely disabled individuals. *IEEE Transactions on Neural System and Rehabilitation Engineering*, *11*(4), 463-469.

KEY TERMS

Morse Code: Morse code is a transmission method, implemented by using just a single switch. The tone ratio (dot to dash) in Morse code has to be 1:3 per definition. This means that the duration of a dash is required to be three times that of a dot. In addition, the silent ratio (dot-space to character-space) also has to be 1:3.

Adaptive Signal Processing: Adaptive signal processing is the processing, amplification and interpretation of signals that change over time through a process that adapts to a change in the input signal.

Augmentative and Alternative Communication (AAC): Support for and/or replacement of natural speaking, writing, typing, and telecommunications capabilities that do not fully meet communicator's needs. AAC, a subset of AT (see below), is a field of academic study and clinical practice, combining the expertise of many professions. AAC may include unaided and aided approaches.

Assistive Technology (AT): A generic term for a device that helps a person accomplishes a task. It includes assistive, adaptive and rehabilitative devices, and grants a greater degree of independence people with disabilities by letting them perform tasks they would otherwise be unable of performing.

Simple Message Service (SMS): A service available on digital mobile phones, which permits the sending of simple messages between mobile phones.

Global System for Mobile Communications (GSM): GSM is the most popular standard for global mobile phone communication. Both its signal and speech channels are digital and it is therefore considered a 2nd generation mobile phone system.

Interworking Architectures of 3G and WLAN

Ilias Politis
University of Patras, Greece

Tasos Dagiuklas
Technical Institute of Messolonghi, Greece

Michail Tsagkaropoulos
University of Patras, Greece

Stavros Kotsopoulos
University of Patras, Greece

INTRODUCTION

The complex and demanding communications needs of modern humans led recently to the deployment of the 3G/UMTS mobile data networks and the wireless LANs. The already established GSM/GPRS radio access technology can easily handle the voice and low-rate data traffic such as short messages (SMS); however, it is inadequate for the more challenging real-time multimedia exchanges that require higher data rates and ubiquitous connectivity. The UTRAN radio access technology provides wide area coverage and multimedia services up to 2Mbps, while the recently deployed WLANs offer radio access at hotspots such as offices, shopping areas, homes, and other Internet/intranet-connected networks, with very high data rates up to 54Mbps (IEEE 802.11g). Hence, there is a strong need to integrate WLANs and 3G access technologies, and to develop a heterogeneous network based on an all-IP infrastructure that will be capable to offer ubiquitous and seamless multimedia services at very broadband rates.

The major benefits that drive towards an all-IP based core network are the following (Wisely et al., 2002):

- **Cost Saving on Ownership and Management:** Network operators need to own and manage one single network, instead of multiple.
- **Cost Saving on Transport:** For example, the cost to provide IP transport is lower.
- **Future Proof:** It can be claimed that the future of backbone network, both for voice and data, is IP based. An IP-based network allows smooth interworking with an IP backbone and efficient usage of network resources.
- Smooth integration of heterogeneous wireless access technologies.
- The IP multimedia domain can support different access technologies and greatly assist towards fix/mobile convergence.
- **Capacity Increase:** The capacity enhancement of an IP-based transport network is quicker and cheaper. The same is also true to service capacity, thanks to the distributed nature of the service architecture.
- **Rich Services:** The benefits of VoIP are available for improved and new services, for example, voice/multimedia calls can be integrated with other services, providing a powerful and flexible platform for service creation.
- Enable peer-to-peer networking and service model.

This hybrid network architecture would allow the user to benefit from the high throughput IP-connectivity in ‘hotspots’ and to attain service roaming across heterogeneous radio access technologies such as IEEE 802.11, HiperLan/2, UTRAN, and GERAN. The IP-based infrastructure emerges as a key part of next-generation mobile systems since it allows the efficient and cost-effective interworking between the overlay networks for seamless provisioning of current and future applications and services (De Vriendt et al., 2002). Furthermore, IP performs as an adhesive, which provides global connectivity, mobility among networks, and a common platform for service provisioning across different types of access networks (Dagiuklas et al., 2002). The development of an all-IP interworking architecture, also referred to as fourth-generation (4G) mobile data network, requires specification and analysis of many technical challenges and functions, including seamless mobility and vertical handovers between WLAN and 3G radio technologies, security, authentication and subscriber administration, consolidated accounting and billing, QoS, and service provisioning (Tafazolli, 2005).

This article discusses the motivation, interworking requirements, and different architectures regarding 3G/WLAN interworking towards an all-IP hybrid networking environment. Five common interworking techniques and architectures that effectively can support most of the issues addressed previously are presented and discussed. These are

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

namely: open coupling, loose coupling, tight coupling, very tight coupling (3GPP, 2004), and the recently developed interworking technology named unlicensed mobile access (UMA), which arises as a very competitive solution for the interworking environment (3GPP-UMAC, 2005). The focus of the article is on a comparison and qualitative analysis of the above architectures.

3G AND WLAN INTERWORKING

Motivation

The main motivation for mobile operators to get involved in the WLAN business (Dagiuklas & Velentzas, 2003) is the following:

- Public WLANs provide the opportunity for mobile operators to increase their revenues significantly from mobile data traffic.
- WLANs can be considered as an environment for testing new applications at the initial stage.
- High-demand data traffic from hotspot areas can be diverted from 3G to WLAN, relieving potential network congestion.
- Location-based services in hotspot areas could be based on WLAN technology rather than using more-complex GPS-like systems.

On the other hand, a shift from WLAN to 3G could take place due to the following reasons:

- **Poor Coverage:** Users may be able to use WLAN services at the airport of departure, but not at the airport of arrival or at the hotel.
- **Lack of Brand Recognition:** The service operators are often new start-ups, which causes end-users to hesitate to use the service.
- **Lack of Roaming Agreements:** End users are forced to locate different service providers at the places they roam to.

The service provider value proposition for utilizing integrated WLANs with cellular networks includes the following benefits for carriers as well as their subscribers:

- Extension of current service offering by:
  - integrating cellular data and WLAN solutions,
  - positioning for voice phone service in hotspots, and
  - engaging enterprises with in-building solutions.
- Improve bottom line with new revenue and lower churn:
  - The carrier provides improved in-building coverage by using intranet bandwidth instead of in-building cell sites to provide coverage.
  - Cross system/service integration features become a competitive advantage for the carriers offering seamless mobility services.
  - The cellular provider derives service revenue for authentication services, mobility services, and calls that do not use cellular bearer channels.
  - The cellular handset becomes an indispensable element.
  - The handset can operate with more functionality, for example, even as gateway.
  - The subscriber increases his dependency on the handset.
- Payload traffic trade-off:
  - Some calls will hand over from cellular channels to WLAN connections when subscribers enter these coverage areas.
  - Other calls will hand over to cellular bearer channels when people leave WLAN coverage areas.
  - A more integrated approach to data traffic will probably increase the use of data transferred over cellular networks.

It becomes evident that as subscribers become more dependent on their much more useful handsets, they will call and be called more and everywhere.

3G and WLAN Architectures

The interworking between 3G and WLAN is a trivial issue that is under study by international standardization fora, namely, ETSI, 3GPP, and the UMTS Forum. The undergoing investigation has provided specific requirements that interworking solutions need to meet. The demands include the establishment of some kind of partnership between the 3G operator and the wireless Internet service provider (WISP), a common billing and accounting policy between roaming partners, and a shared subscriber database for authentication, authorization, and accounting (AAA) and security provisioning (Nakhjiri & Nakhjiri, 2005).

The work in this article refers to four already established interworking scenarios (Salkintzis, 2004) regarding 3G and WLAN, which are presented and compared to the recently developed UMA architecture.

In *open coupling* interworking architecture, there is no requirement for specific WLAN access, while each of the networks—3G and WLAN—follows separate authentication procedures. This architecture does not support seamless services, while the user performs a vertical handover from 3G towards WLAN and vice versa.

Figure 1. Open coupling scenario

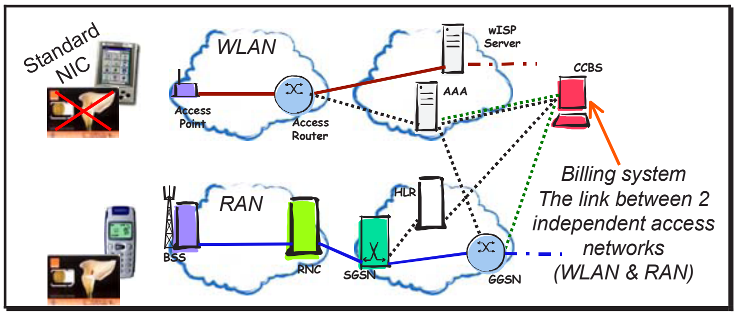

Standard NIC
WLAN
wISP Server
AAA
CCBS
Access Point
Access Router
RAN
HLR
BSS
RNC
SGSN
GGSN
Billing system
The link between 2 independent access networks (WLAN & RAN)

Figure 2. Loose coupling scenario

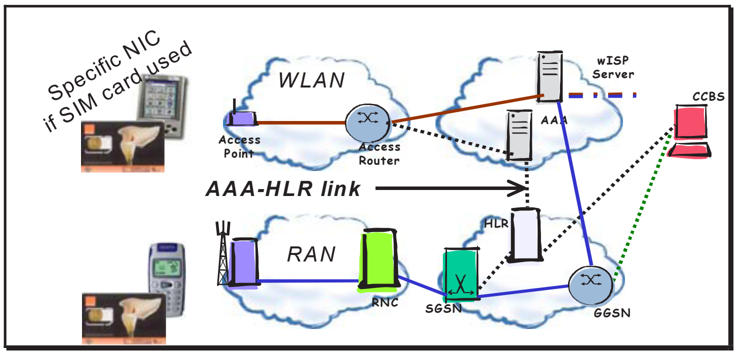
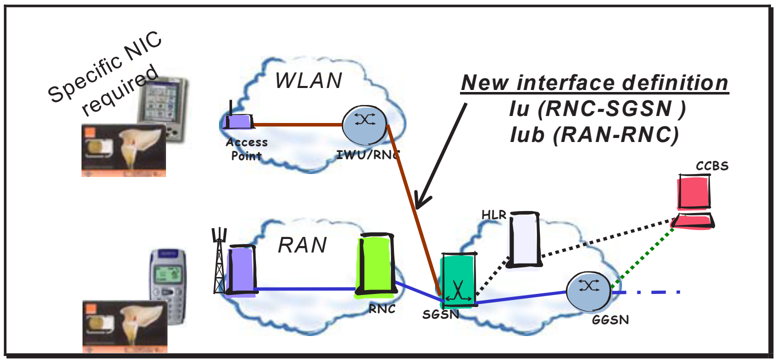

Specific NIC if SIM card used
WLAN
wISP Server
CCBS
AAA
Access Point
Access Router
AAA-HLR link
HLR
RAN
RNC
SGSN
GGSN

Figure 3. Tight coupling scenario


Specific NIC required
WLAN
Access Point
IWU/RNC
New interface definition
Iu (RNC-SGSN)
Iub (RAN-RNC)
CCBS
HLR
RAN
RNC
SGSN
GGSN

In a *loose coupling* scenario, 3G and WLAN share a common customer database and authentication process. There is no requirement for specific WLAN access. In addition, no load balancing is provided for applications with specific QoS requirements, and the architecture does not support seamless services. Similar to previous architecture, loose coupling does not support seamless services.

On the other hand, a *tight coupling* scenario supports seamless service provisioning for vertical handovers and load balancing for QoS demanding applications. However, there is need for definition of the interface interconnection between the WLAN and SGSN node.

Similar to the previous scenario, *very tight coupling* offers the same advantages, although in this case WLAN is considered as part of the UTRAN and a new interface interconnecting the WLAN and the UTRAN/RNC needs specification.

Figure 4. Very tight coupling scenario

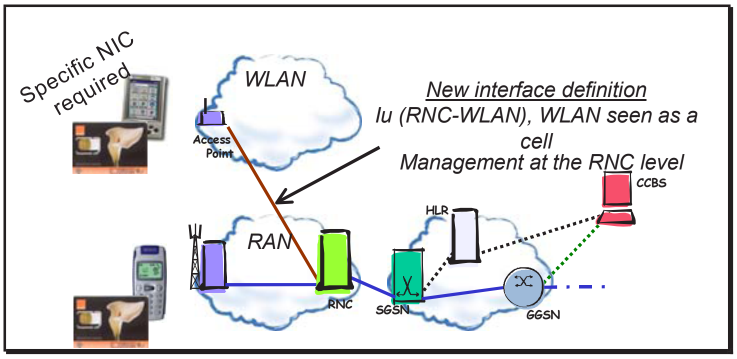
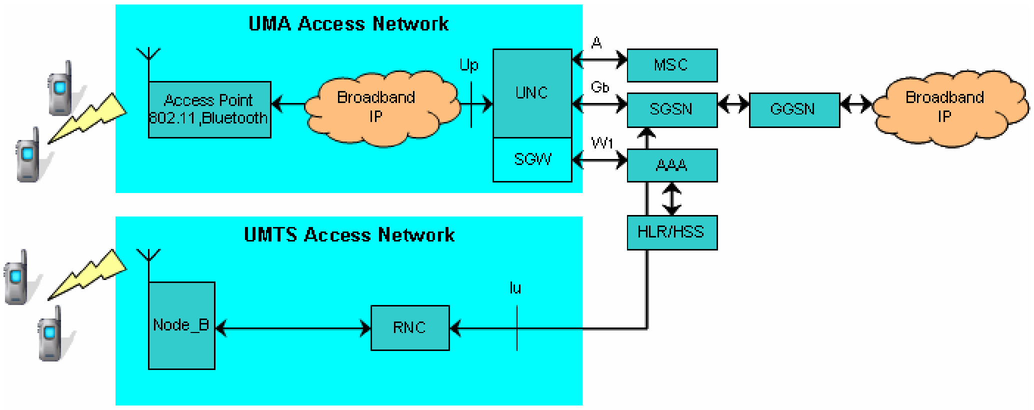

Specific NIC required
WLAN
Access Point
New interface definition
Iu (RNC-WLAN), WLAN seen as a cell
Management at the RNC level
CCBS
HLR
RAN
RNC
SGSN
GGSN

Figure 5. UMA architecture


UMA Access Network
Access Point
802.11,Bluetooth
Broadband IP
Up
UNC
SGW
A
Gb
W1
MSC
SGSN
GGSN
Broadband IP
AAA
HLR/HSS
UMTS Access Network
Node_B
RNC
Iu

The above interworking scenarios are constrained by the fact that they cover only three wireless technologies. The B3G vision considers the employment of new emerging wireless technologies (e.g., WiMAX, 4G, etc.). The limitation of the approach mentioned above can be alleviated through the approach envisioned by the Unlicensed Mobile Access Consortium (UMAC), where all wireless technologies can be smoothly integrated towards an all-IP-based heterogeneous network. Recently, UMAC produced a multi-access architecture known as unlicensed mobile access (UMA), which rises as an exiting prospect. In the next section, the UMA architecture is presented.

UNLICENSED MOBILE ACCESS

The UMA technology provides access to circuit and packet-switched services over the unlicensed spectrum, through technologies which include Bluetooth and IEEE 802.11 WLANs. The standardization procedure has been initiated by the Unlicensed Mobile Access Consortium (UMAC) and is currently continued by the 3rd Generation Partnership Project (3GPP) under the work item "Generic Access to A./Gb interface." UMA deployment offers to the subscribers a ubiquitous connectivity and consistent experience for mobile and data services as they transition between cellular and public or private unlicensed wireless networks. Compared to other cellular-WLAN interworking solutions, UMA is superior in both the technology aspects and the cost effectiveness for customers and providers (Velentzas & Dagiuklas, 2005).

In more detail, UMA technology does not affect the operation of current cellular radio access networks (UTRAN/GERAN), like cell planning. Hence the investments in existing and future mobile core network infrastructures are preserved. UMA utilizes standard all-IP infrastructure, which provides access to standard packet-switched services and applications. Additionally, the UMA network (UMAN) is independent of the technology used for unlicensed spectrum access, thus it is open to new wireless technologies with no extra requirements. It enables seamless handover between the heterogeneous access technologies, while it ensures service continuity across the network coverage area. Compared to a loose-coupling WLAN-3G interworking scenario, UMA offers greater control over the authentication, authorization, and accounting procedures and tighter end-to-end security. Moreover, UMA technology supports load balancing for efficient allocation of bandwidth and data rates according to the customer requirements, and provides higher throughput and network capacity to the operator that it is translated to more connected customers, thus producing higher revenues.

The UMA technology introduces a new network element, the UMA Controller (UNC), and associates protocols that provide secure transport of GPRS signaling and user plane traffic over IP. UNC acts as a GERAN base station controller (BSC) and includes a security gateway (SGW) that: (1) terminates secure remote access tunnels between the UNC and the mobile station (MS); and (2) provides authentication, encryption, and data integrity for signaling and media traffic. The UNC is connected to a unique mobile switching center (MSC) through the standard A-interface and a serving GPRS support node (SGSN) through the Gb-interface in order to relay GSM and GPRS services respectively to the core network, while through the Wm-interface allows authentication signaling with the corresponding AAA server. The Up-interface between the UNC and MS operates over the IP transport network, and relays circuit and packet-switched services and signaling among the mobile core network and the MS. The MS, which is capable of switching between cellular radio access networks (RANs) and unlicensed, also has an IP interface to the access point that extends the IP access from the UNC to the MS.

COMPARISON OF DIFFERENT ARCHITECTURES AND QUALITATIVE ANALYSIS

The following table showcases the comparison between the proposed interworking architectures. The comparison is based on already mentioned parameters that affect not only the technical efficiency of each of the proposed solutions, but also the cost efficiency, as this is perceived from the operator's and from the user's point of view.

Table 1. Qualitative comparison of interworking scenarios

		Open Coupling	Loose Coupling	Tight Coupling	Very Tight Coupling	UMA
Qualitative Parameters	**Service Continuity**	The running application will not continue across 3G and WLAN once vertical handover takes place	It is not supported and time sensitive services will be interrupted during handover	Service continuity is supported, although QoS may be degraded during handover	Similar to Tight Coupling	Service continuity is fully supported
	Simplicity	The user for the same service may have to subscribe to at least two service providers	The user for the same service may have to subscribe to at least two service providers	One service, one mailbox	One service, one mailbox	One service, one mailbox
	Seamlessness	Seamless vertical handovers are not supported	Seamless services are not supported by this architecture	Seamless handovers and mobility are supported	Seamless handovers and mobility are supported	Seamless mobility for circuit and packet switched services
	Load Balancing	The architecture has no capability to divert the services according to their QoS demands	It is no supported, the system cannot select the network that is suitable for the QoS requests of the service	It is not supported, network selection is based on network coverage at current user's location	It is not supported, network selection is based on network coverage at current user's location	Supports load balancing decisions based on the required QoS of the application and the application's requested bandwidth
	Security and Authentication	Separate authentication procedure and security provisioning	Common authentication procedure, the 3G/HLR database is shared between the networks	Common authentication procedures, SGSN is the point of decision, 3G like security scheme	Common authentication procedures and security provisioning	The architecture supports SIM/EAP-AKA authentication and IPSec protocol for the unlicensed mobile part and common GPRS/3G authentication procedure from UNC towards SGSN and MSC

Table 1. Qualitative comparison of interworking scenarios, continued

Qualitative Parameters	**Openness**	Additional radio access schemes may be added to the architecture	The architecture may support other heterogeneous wireless technologies	The architecture is proprietary and depends on the WLAN technology used	The architecture is proprietary and depends on the WLAN technology used	The architecture supports interworking between 3G and heterogeneous wireless technologies
	Access Control	No specific WLAN access is required, the access control is WLAN based	Independent of access technology used, the access control is WLAN based	Dependent on access technology used due to the new IWU/RNC-SGSN interface, access control is 3G based	Dependent on access technology, management at RNC, 3G based access control	Access is WLAN based over unlicensed radio spectrum
	MT Complexity	Standard MT are used with some kind of WLAN interface	Standard MT are used with some kind of WLAN interface	MT with specific NIC	MT with specific NIC	MT are required to support mechanism (software module) for switching between UMA and 3G radio interface
	Standardization Complexity	No further standardization effort is required	ISP/AAA-3G/HLR link standardization, translation from MAP to RADIUS/DIAMETER	Significant standardization effort, new interface definition, 'Iu' (RNC-SGSN), 'Iub' (RAN-RNC)	'Iu' interface between RNC and WLAN	Further standardization is required, however leverages many already defined 3G and IP based protocols
	Cost Efficiency	WLAN and 3G networks run separately and there is no further financial burden for the operator	A very cost efficient solution since it is based on the implementation of well established technologies	The operator is required to install new infrastructure at hotspots to interconnect WLAN in the SGSN	The operator is required to install new infrastructure at hotspots to interconnect WLAN in the RNC	New infrastructure needs to be installed, UNC, however utilisation of unlicensed spectrum and openness to new technologies compensates for the cost.

In comparison with the loose and tight coupling scenarios, the UMA approach for WLAN 3G interworking has the advantage in both technology and cost efficiency for operators and customers. Specifically, the installation of a UNC controller, although it requires further standardization and protocol definitions, arises as a better solution. It allows easy openness to new wireless technologies with no extra requirements and is able to provide QoS guarantees for multimedia real-time applications and time-sensitive services in general, seamless mobility, and uninterrupted services across the network coverage area. Hence, the operator is able to cover the demands of customers and the customers have access to all services anywhere and anytime. Authentication, authorization, accounting, and security are common between 3G and WLAN and based on already established protocols and standards. Compared to the other solutions, UMA offers greater control on AAA procedures and tighter end-to-end security. Finally, UNC supports load balancing, which provides—in addition to the best possible solution in terms of connectivity, bandwidth, and data rates, according to the service that the customer requires—higher throughput and network capacity to the operator so that it is translated to more connected customers, thus producing higher revenues.

FUTURE TRENDS

There is no industry consensus on what next generation networks will look like, but as far as the next generation networks are concerned (Kingston, Morita, & Towle, 2005), ideas and concepts include:

- transition to an "All-IP" network infrastructure;
- support of heterogeneous access technologies (e.g., UTRAN, WLANs, WiMAX, xDSL, etc.);
- VoIP substitution of the pure voice circuit switching;
- seamless handovers across both homogeneous and heterogeneous wireless technologies;
- mobility, nomadicity, and QoS support on or above IP layer;
- need to provide triple-play services creating a service bundle of unifying video, voice, and Internet;
- home networks are opening new doors to the telecommunication sector and network providers;
- unified control architecture to manage application and services; and
- convergence among network and services.

Two important factors have been considered to satisfy all these requirements. The first one regards the interworking of existing and emerging access network under the umbrella of a unified IP-based core network and unified control architecture supporting multimedia services. A proposed solution towards this direction is the unlicensed mobile access (UMA), allowing heterogeneous wireless technologies to interconnect to a core network through a network controller. The second requirement regards IP multimedia subsystem (IMS) evolution in order to cope with requirements imposed by NGN architecture (Passas & Salkintzis, 2005). The initial release of 3GPP IMS was developed only for mobile networks. The increasing demand of interworking between different access devices and technologies led to subsequent releases that defined IMS as a core independent element and a key enabler for applying fixed mobile convergence (FMC). FMC comprises two attributes: using one number, voice/mail and seamless handover of multimedia sessions. In the B3G/4G vision, IMS is required to become the common architecture for both fixed and mobile services. Towards this end the ETSI Telecoms and Internet converged services and protocols for advanced networks (TISPAN) is also producing new functionality extensions for the IMS (ETSI TISPAN, n.d.).

CONCLUSION

The conclusion of the qualitative analysis relates the most suitable solution for an interworking architecture of 3G and WLAN radio access technologies based on an all-IP core network with the UMA network. The most important characteristics of UMA, as they have been discovered during the analysis, are among others: the seamless support for vertical handovers and the QoS guarantees for multimedia and time-sensitive applications due to the load balance capability; and the network continuity, scalability, and cost efficient openness. Moreover, the UMA network solution for integrated 3G and WLAN technologies enables network operators to leverage cost and performance benefits of VoIP, broadband, and Wi-Fi, while it supports all mobile services voice, packet, and IMS/SIP, and utilizes standard interfaces into the all-IP core network.

REFERENCES

Dagiuklas, T. et al. (2002). Seamless multimedia services over all-IP network infrastructures: The EVOLUTE approach. *Proceedings of the IST Summit 2002* (pp. 75-78).

Dagiuklas, T., & Velentzas, S. (2003, July). *3G and WLAN interworking scenarios: Qualitative analysis and business models.* IFIP HET-NET03, Bradford, UK.

De Vriendt, J. et al. (2002). Mobile network evolution: A revolution on the move. *IEEE Communications Magazine, 4,* 104-111.

ETSI TISPAN. (n.d.). *NGN functional architecture: Resource and admission control subsystems, release 1.*

Kingston, K., Morita, N., & Towle, T. (2005). NGN architecture: Generic principles, functional architecture and implementation. *IEEE Communications Magazine,* (October), 49-56.

Nakhjiri, M., & Nakhjiri, M. (2005). *AAA and network security for mobile access* (pp. 1-23). New York: John Wiley & Sons.

Passas, N., & Salkintzis, A. (2005). WLAN/3G integration for next generation heterogeneous mobile data networks. *Wireless Communication and Mobile Computing Journal,* (September).

Salkintzis, A. (2004). Interworking techniques and architectures for WLAN/3G integration towards 4G mobile data networks. *IEEE Wireless Communications,* (June), 50-61.

Tafazolli, R. (2005). *Technologies for the wireless future.* New York: John Wiley & Sons.

3GPP. (2004a, September). *Technical specification group services and system aspects: 3GPP system to wireless local area network (WLAN) interworking: System description, 3 TS 23.234.* Retrieved from http://www.3gpp.org

3GPP. (2004b, September). IP multimedia subsystem version 6. 3G TS 22.228.

3GPP-UMAC. (2005, June). *UMA architecture (stage 2).* Retrieved from http://www.3gpp.org

Unified Mobile Access Consortium. (n.d.). Retrieved from http://www.uma.org

Velentzas, S., & Dagiuklas, T. (2005, July). *Tutorial: 4G/wireless LAN interworking.* IFIP HET-NET 2005, Ilkley, UK.

Wisely, D. et al. (2002). *IP for 3G: Networking technologies for mobile communications.* New York: John Wiley & Sons.

KEY TERMS

Authentication Authorization Accounting (AAA): Provides the framework for the construction of a network architecture that protects the network operator and its customers from attacks and inappropriate resource management and loss of revenue.

B3G/4G: Beyond 3G and 4G mobile communications that provide seamless handover between heterogeneous networks and service continuity.

IP Multimedia Subsystem (IMS): Provides a framework for the deployment of both basic calling and enhanced multimedia services over IP core.

NGN: An ITU standard for Next Generation Networks where cellular mobile 3G systems, WLANs, and fixed networks are integrated over IP protocol.

3G: Third generation of cellular mobile communications (GPRS/UMTS).

iPod as a Visitor's Personal Guide

Keyurkumar J. Patel
Box Hill Institute, Australia

Umesh Patel
Box Hill Institute, Australia

INTRODUCTION

Over the past few years, use of mobile devices for various purposes has increased. Apple released its first iPod on October 23, 2001, a breakthrough MP3 player. Today, Apple's fifth-generation iPod is available which can be considered as a portable media player that focuses on the playback of digital video, as well as storing and displaying pictures and video (see apple.com*)*. Since then the iPod has been successfully and effectively used for various purposes including as a media player, bootable drive, external data storage device, PDA replacement, and for podcasting.

Academia and tourism are two areas where the use of mobile devices are encouraged to gain benefits from the technology. For academic use, the iPod's recording and storage capabilities have been explored by some educational institutes across the United States. According to the Duke University iPod First-Year Experience Final Evaluation Report, the iPod supports individual learning preferences and needs, and easy-to-use tools for recording interviews, field notes, and small-group discussions. The tourism industry is also identified as a potential area to use mobile technologies. Recently, Dublin Tourism, Ireland discovered the use of the iPod as a portable tourist guide; Ireland's neighbor Scotland followed (see Physorg, 2006).

Sales of interactive portable MP3 players have increased explosively in the last few years. Information Media Group predicts that sales will continue to increase at the rate of 45% for next six years (Macworld UK, 2005). The iPod is currently the world's best-selling digital audio player and increased its popularity in Australia sevenfold in 2004 (see apple.com). Greg Joswiak, the worldwide vice president of iPod marketing, said: "As of August 2005, market share in Australia is 68% of [the] digital player market."

With the increasing use of digital media together with the handheld devices, this iPod application will eliminate the need for human guides and will provide an entertaining experience to visitors. It will be very useful for landmark tourist destinations such as aquariums and museums, and there will be a huge demand with the increasing flow of tourists in Australia, which according to Tourism Australia (2005) was an increase of 5.4% from 2004 to 2005, with tourists numbering 5.5 million in the latter year.

BACKGROUND

Tourism is an important activity for human life as a source of pleasure and during the holidays. We visit various places every now and then, including tourist destinations such as a museum, commercial destinations such as a stock market, educational institutes such as a university, or public places such as a shopping mall.

Every new visitor suffers from preconceptions and anxiety from their lack of knowledge about the visiting site. This acts as a barrier and must be overcome before an effective visit can take place. As for visitors' preconceptions, the authors of this article encourage visitors to address their anxiety and introduce excitement before they start the tour. The tourism industry so far has promoted the various communication mediums such as maps on the board, written information about specific locations, and now display video screens. Tourism has been a popular area for mobile information systems (Cheverst, Davies, Mitchell, Friday, & Efstratiou, 2000) and other PDA-based systems.

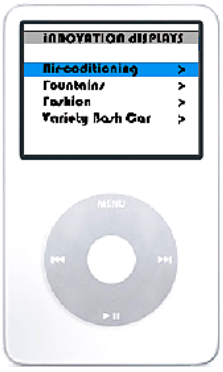
Audio and video has been neglected or underused as a leaning medium in recent years (Scottish Council for Educational Technology, 1994.). The general view is that video is a better tool for leaning than audio. Animation and interactive media like simulations can attract attention, but they proved to be expensive. Hearing is an astoundingly efficient skill according to Clark and Walsh (2004).

Portable media players such as PDAs and iPods can provide information anytime and anywhere. These devices come with their own hard drives and eliminate transportation of storage devices, which is a requirement for video communication. The iPod, with built-in speakers and microphone, makes it easier to record and playback information stored into its hard drive. Clark and Walsh (2004) stated that besides its popularity and ease of use, listening to an iPod and similar devices at public places is socially accepted.

At Box Hill Institute of TAFE, we realized the use of an iPod as a part of our "Innovation Walk" project. The "Innovation Walk" is developed with the aim of showcasing the institute's prized innovations. Career teachers, overseas visitors, students, industry and government dignitaries, and member of the community can undertake the walk independently or as a guided tour.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Designed main menu of a prototype device



A prototype device is being developed using an Apple iPod (see Figure 1).

PERSONAL GUIDE DESIGN

The visitor's personal guide itself will be in the form of an iPod, which can be programmed to give details of a defined list of locations, as well as playing an audible narration of each featured location. This will allow visitors to navigate the visitor's site on their own with the use of the iPod. The following technologies were considered initially to program the iPod as per the requirements:

- creating an application in J2ME on the Java platform porting it to the iPod;
- installing a variant of Linux (more on this later), and modifying its operation to create the system from this platform; and
- creating a text-based guide on the iPod.

The text-based option is the easiest way, with some limitations and given preference on the basis of the estimated time and skills available. To get multiple text pages to run is a fairly simple concept. It requires a specifically named file located under the "Notes" folder that acts as an index page, from which the menu would be created and all other notes will be created. As discussed earlier, iPods as storage devices can easily be connected to a computer via the USB port, and the drive that is mounted for the iPod can be navigated easily from "My Computer."

Creating a Content Page

Open up the "Notes" folder and create a new file called "Main.linx." This file name is required for two reasons. The first reason is that by naming the file "Main," the iPod will automatically display the page as an index, rather than providing a list of available files to open. Secondly, the extension ".linx" of the filename defines the method used to display on the iPod screen a link to another text file.

The iPod has two methods of displaying a link. The default is to have a link created within a text file appear as a hyperlink, similar to that of an html Web page with the word or sentence underlined. The second method is to display the link as an actual menu item on the iPod. This method would be ideal for the contents of our visitor's guide.

Once the "Main.linx" file has been created and located correctly, the next step will be the contents of the main page. This will create the major links to each of the locations that will contain information. This is achieved by opening up the "Main.linx" file in the Notepad and entering the following:

```
<title>
```

Alternative Operating Systems for Apple's iPod

Currently there are two main alternatives to Apples' iPod Operating System: iPodLinux (an open source venture into porting Linux onto the iPod) and Rockbox (an open source replacement firmware for mp3 players).

iPodlinux (www.ipodlinux.org) and Rockbox Operating System (www.rockbox.org) are able to replace Apple's Operating System and still maintain the same functionality. The alternative operating systems are capable of playing mp3s and other audio formats, videos, and reading notes. The main difference between Apple's Operating System is that with iPodlinux and Rockbox you can:

- play video games,
- run applications,
- simply develop your own applications without requiring commercial development tools, and
- programmers can develop their own applications or modify/customize existing iPodlinux GPL (General Public Licenxe) applications.

Certain Linux applications are recompiled to run on the iPod without modification. Both alternatives to Apple's iPod Operating System have a following of enthusiastic programmers and developers who have figured out the workings of the five generations of the iPod. Developers and programmers of the iPodlinux have contributed a lot to an open source operating system by setting up Internet relay chat rooms, news groups, forums, wikis, and Web sites. Sourceforge hosts the source code and development comma separated value (CSV) tree, which is maintained by the iPodlinux core developers. Documentation of the iPod hardware components such as the microcontroller, display, memory, battery, and so on is now accessible to everyone. Rockbox Operating System developers thank the hard work of the iPodlinux project because if it was not for iPodlinux documentation and developers, the Rockbox Operating System may have never been ported to the iPod.

Why Choose an Alternative iPod Operating System

The iPod as a visitor's personal guide project initially was looking at the bleeding edge mobile Java J2ME application technology to fulfill its requirements. After research it was discovered that there are other ways to implement a tour guide on an iPod. The research found iPod Notes, iPodlinux, and Rockbox.

- The iPod Apple Operating System is proprietary and therefore a close source.
- iPodLinux and Rockbox are open source operating systems written under the GNU General Public License.
- iPod Apple OS only supports a crippled html language in "Notes" which allows the development of interactive Notes that can contain pictures, video, and text.

How iPod Can be Programmed for iPodLinux

Programming for iPodLinux is done in C, and as a prerequisite the standard functions and libraries must be used. Here is an example of the "Hello World" code using the print function from the stdio.h library.

Using notepad or a C application, do the following:

- Start off by including the precompiler derivative includes statement: #include <stdio.h>.
- Next create the main function from which we will put in the code to print Hello World (see Figure 2).
- Now save the code you have entered into the notepad using the filename of hworld.c. The step is to compile hworld.c using the arm complier tools, arm-elf-gcc hworld.c -o hworld -elf2flt.
- Executing hworld on the iPod running iPodlinux will display "Hello World" to the iPod screen. Once you are happy with your application, it can be packaged as a module and inserted into the Podzilla menu structure.

Figure 2. Sample main function created using programming language "C"

```
int main (int argc, char **argv)
{
printf ("Hello World!\n" );
return 0;
}
```

ADVANTAGES

This new application and use of iPod will:

- eliminate the need for a human guide;
- provide a self-guided tour with entertainment to visitors;
- lead to interactive customer service;
- provide flexibility to tourists to tour the area per their own need, time, and interest, which is an important perspective; and
- achieve great tourist turnaround, as there is no need to wait for some predefined number of tourists.

FUTURE TRENDS

This concept can further be considered to provide visitor information in other languages than English, with possible navigation for the use of different languages in a multicultural environment. Also our next application will discover the possibilities of porting iPodLinux platform on the fifth-generation iPod which is not done so far. Further, the possibility of using an iPod—similar to a PDA—in a commercial environment will be investigated. For now, this cost-effective solution can be implemented at various landmark tourist destinations such as mines, aquariums, and museums, and in the near future it will replace existing expensive technologies.

CONCLUSION

We have demonstrated that the iPod can be used as an innovative and cost-effective tool. To realize the use of the iPod as a visitor's personal guide, iPod's simple user interface designed around a central scroll wheel can be explored for the navigation and recording/ playback facility. It provides the latest information to visitors. Furthermore, iTunes can add an entertaining experience with preferable music while using it as a personal guide.

ACKNOWLEDGMENTS

We would like to thank our final-year students Andrew Obersnel, Ben Coster, Randima Sampath Ratnayake, and Pathum Wickasitha Thamawitage for their contributions toward this project. We would also like to thank Rob McAllister,

general manager of teaching, innovation, and degrees, Box Hill Institute; Stephen Besford, center manager, Center for Computer Technology, Box Hill Institute; and John Couch, administrative officer, Center for Computer Technology, Box Hill Institute.

REFERENCES

Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C. (2000). Developing a context-aware electronic tourist guide: Some issues and experiences. *Proceedings of CHI 2000* (pp. 17-24). The Hague: ACM Press.

Clark, D., & Walsh, S. (2004). *iPod learning.* White Paper, Epic Group, Brighton, UK.

Duke University. (2005). *Duke iPod first year experience.* Retrieved February 17, 2006, from http://www.duke.edu/ipod/

Macworld UK. (2005). Music player sales 'set to double'. *Macworld UK,* (July 22). Retrieved April 26, 2006, from http://www.macworld.co.uk/news/index.cfm?NewsID=9218&pagePos=5

Physorg. (2006). *Portable tourist guide now on service.* Retrieved February 19, 2006, from http://www.physorg.com/news10338.html

Scottish Council for Educational Technology. (1994). Audio. In *Technologies and learning* (pp. 24-25). Glasgow: SCET.

Tourism Australia. (2005). *Visitors arrival data.* Retrieved February 19, 2006, from http://www.tourism.australia.com/Research.asp?sub=0318&al=2100

KEY TERMS

Java2 Macro Edition (J2ME): A Java platform especially for programming mobile devices such as PDAs.

Linux: An open source operating system for computers.

MP3: MPEG-1 Audio Layer 3, of sound or music recordings stored in the MP3 format on computers.

Personal Digital Assistant (PDA): A mobile device that can be used both as a mobile phone and a personal organizer primarily.

Universal Serial Bus (USB) Port: Port used to connect devices to computers such as PCs, laptops, and Apple Macintosh computers.

Keyword-Based Language for Mobile Phones Query Services

Ziyad Tariq Abdul-Mehdi
Multimedia University, Malaysia

Hussein M. Azia Basi
Multimedia University, Malaysia

INTRODUCTION

A mobile system is a communications network in which at least one of the constituent entities—that is, user, switches, or a combination of both—changes location relative to another. With the advancements in wireless technology, mobile communication is growing rapidly. There are certain aspects exuded by mobile phones, which make them a high potential device for mobile business transactions. Firstly, there is a growing statistic on the number of users who own at least one mobile phone. In 2003 alone, the numbers of mobile phone users were as high as 1.3 billion, and this number is growing steadily each year. Secondly, more and more mobile phones are equipped with much better features and resources at a considerably lower price, which make them affordable to a larger number of users. And thirdly, and most importantly, the small size of mobile phones makes them easily transportable and can truly be the device for anywhere and anytime access (Myers & Beigl, 2003).

Database querying, which is the interest of this article, is a kind of business transaction that can benefit from mobile phones. In general, database querying concerns the retrieval of information from stored data (kept in a database) based on the query (request) posed by the users. This aspect of the database transactions had been the focus of many database researchers for a long time. The mobile phone aspect of the transaction had only recently gained interest from the database communities, and these interests were mainly targeted to the "fatter" mobile devices. The work on mobile database querying can be grouped into those focused on the technology and techniques to handle the limitations of the mobile transmissions due to the instability of the mobile cellular networks, which were concentrated on developing applications that involved access to databases for the mobile environment, and those that handled both of the above issues. For example, caching (Cao, 2002) and batching (Tan & Ooi, 1998) are some of the popular techniques that were and still are investigated in detail to handle the problems of the mobile transmissions. On the other hand, Hung and Zhang's (2003) telemedicine application, Koyama, Takayama, Barolli, Cheng, and Kamibayashi's (2002) education application, and Kobayashi and Paungma's (2002) Boonsrimuang transportation application are some examples of the work on mobile database application. These works were observed to be application-specific and supported a very limited and predefined number and type of possible queries. Each of the possible queries, in turn, requires several interactions with the server before a full query can be composed.

This article will highlight the framework opted by the authors in developing a database query system for usage on mobile phones. As the development work is still in progress, the authors will share some of the approaches taken in developing a prototype for a relationally complete database query language. This work concentrates on developing an application-independent, relationally complete database query language. The remainder of this article is organized as follows. The next section presents some of the existing work related to the study. We then introduce and describe the framework undertaken in order to develop a database query system for mobile phones, and discuss the prototype of the database query language used by the system. We end with our conclusion.

RELATED WORK

Query languages are specialized languages for asking questions, or queries, which involve data in a database (Ramakrishnan & Gehrke, 2000). Query languages for relational databases originated in the 1970s with the introduction of relational algebra and relational calculus by E.F. Codd. Both relations are equivalent in their level of expressiveness or query completeness. These two formal relations had interchangeably been used as the benchmarks for measuring the completeness of the later query languages. Codd originally proposed eight operations to be included in the relational algebra, but out of the eight, five were considered fundamentals as they allowed most of the data retrieval operations. These operations are known as selection, projection, cartesian product, union, and set difference. If a query language supports the five operations, then it is considered as being relationally complete (Connolly, Begg, & Strachan, 1997). Throughout

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the years, several other measures of query completeness were proposed such as datalog, stratified, computable, and others (Chandra, 1988). However, in the authors' opinion, these later measures might be too extensive to be considered for mobile phones and their users' application.

Although both relational algebra and relational calculus are complete, they are hard to understand and use. This resulted in a search for other easy-to-use languages that are at least compatible to relational algebra and calculus. Some of these query languages are transform-oriented non-procedural-based languages, which use relations to transform input data into required outputs. Structured Query Language (SQL) is an example of such a language. Besides non-procedural languages, visual query languages have also gained much acceptance in the database community. Some of the work on visual query languages found in the literature, such as Czejdo, Rusinkiewicz, Embley, and Reddy (1989), Catarci (1991), and Polyviou, Samaras, and Evripidou (2005), used the entity relationship diagram and other data modeling as the basis for query formulation, and some used icons to present pre-defined queries (Massari, Weissman, & Chrysanthis, 1996). Query languages are textual languages that caught the interest of some database query language researchers. Some of these languages were represented in the form of natural language (Kang, Bae, & Lee, 2002; Hongchi, Shang, & Ren, 2001), and some were represented in the form of keywords (Calado, da Silva, Laender, Ribeiro-Neto, & Vieira, 2004; Agrawal, Chaudhuri, & Das, 2001). This type of languages is less restrictive compared to the other types of languages. However, they need extra work in approximating the meaning of the terms or keywords used in a query. Thesaurus and ontology are few approaches used to approximate meanings of terms or keywords (Kimoto & Iwadera, 1991; Weibenberg, Voisard, & Gartmann, 2006) in this type of query language.

Even though each type of query language mentioned above has its own advantages, very few of them, except for Polyviou et al. (2005) and Massari et al. (1996), were tested on small devices. SQL, for example, would be too tedious to be entered using mobile phones and too complicated for ordinary users. Visual query languages, on the other hand, would require considerable screen space as well as resource consuming in order to be rendered. Natural language and keyword language would also be difficult to be textually keyed in using mobile phones. There were, however, attempts to use spoken method for query languages (Chang et al., 2002; Bai, Chen, Chien, & Lee, 1998). But this approach leads to another problem in matching the intonation of users. The textual form of query languages (keyword method in particular) might be the most suitable language to be used on mobile phones since they are the least resource consuming and easily extensible. However, there must be a method to ease the input part of the query formulation process. To date, the authors have not been able to find any publication of the investigations of the same method as applied to mobile phones. Therefore, we believe that the keyword-based language is worth some investigation.

Framework Model

Polyviou, Samaras, and Evripiou (Kang et al., 2002) laid down several challenges that must be dealt with in order to develop a modern search interface. The challenges specified were: the search interface must be usable, powerful, flexible, and scalable. These challenges are adopted in our approach while developing the database query system for mobile phones. The concept of usable is implemented in our design by providing a language that supports menu-based guidance for the users to form valid queries. The concept of powerful is implemented by making sure that the language is relationally complete. The concept of flexible, on the other hand, is implemented in the language by allowing the language to work with any type of relational databases and any type of applications. Finally, the scalability aspect is handled by allowing the language to accept a database of any size, but at the same time filtering the data to be presented to the users according to some form of user grouping and access patterns. The keyword-based language is developed to answer the above challenges. The reasons for choosing such a language, among others, are due to the ability of such a language to present complex relationships with a minimal number of keywords, and it takes lesser space for presentation. For example, it is possible to access information from two indirectly linked relations, no matter how far apart the relations might be, by simply providing the name of the two relations as query keywords. This ability makes the language scalable and easily extensible. However, keyword-based language does have constraints. Firstly, it is in textual form and therefore is tedious and prone to typing error. Secondly, it requires users to have exact knowledge of the keywords to be entered in order to form valid queries. Therefore, the authors have modified the keyword-based approach by providing users with selectable keywords in a menu form. This approach has another advantage: it allows users to point and click the keywords needed without having to type them manually, which is a way to handle the input mechanism problem of mobile phones (most phones only have keypads as an input mechanism). This approach requires lengthy display space, which is lacking for mobile phones. Therefore, the authors intend to handle this problem by providing only selected keywords to users based on their personal profiles and preferred queries. Figure 1 shows the general framework of the query language, and Figure 2 shows the position of the query language with respect to the rest of the whole database query system. As shown in both figures, the query language basically resides in two locations: in the mobile phones as the query interface, and in the application server as the query engine that transformed the keyword

Figure 1. General framework for query language

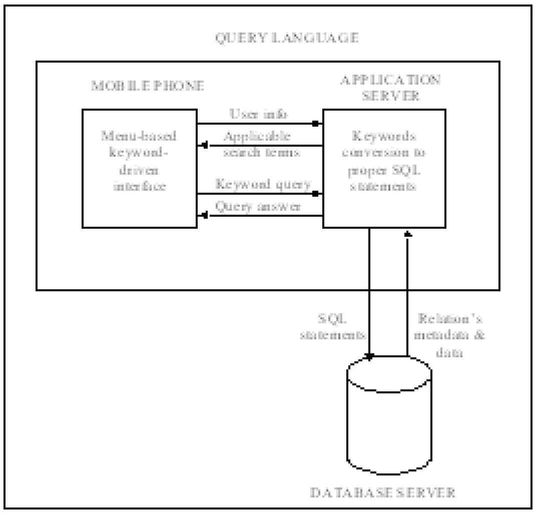

QUERY LANGUAGE
MOBILE PHONE
APPLICATION SERVER
Menu-based keyword-driven interface
User info
Applicable search terms
Keyword query
Query answer
Keywords conversion to proper SQL statements
SQL statements
Relation's metadata & data
DATABASE SERVER

Figure 3. Activities during pre-language operational stage

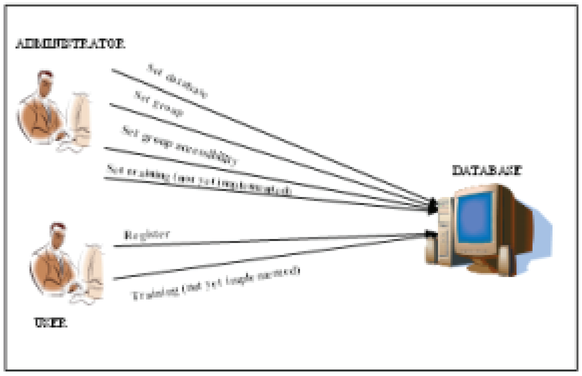

ADMINISTRATOR
Set database
Set group
Set group accessibility
Set training (not yet implemented)
DATABASE
Register
Training (not yet implemented)
USER

Figure 2. The position for database query system

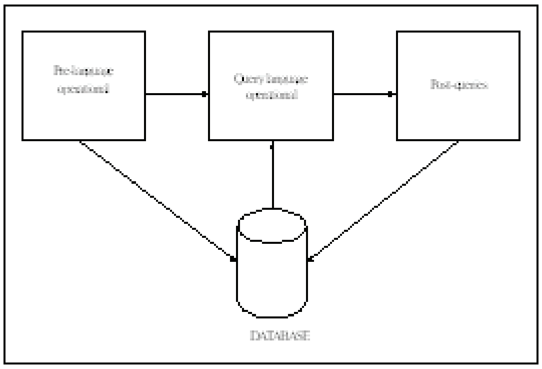
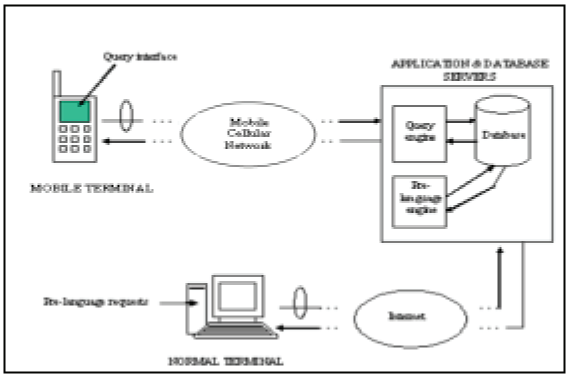

Pre-language operational
Query language operational
Post-queries
DATABASE

Figure 4. The operational stage of network architecture


Query interface
MOBILE TERMINAL
Mobile Cellular Network
APPLICATION & DATABASE SERVERS
Query engine
Database
Pre-language engine
Pre-language requests
Internet
NORMAL TERMINAL

queries into their relevant Structured Query Language (SQL) queries. The keywords that are presented as options during query formulation are selected metadata (i.e., relations' names and attributes' names) of the relations in the database. The selected metadata is based on the user information provided during login (currently personal information; later we will consider results from group training as well) and the accessibility information set by the database administrator prior to the query language becoming operational.

Figure 3 shows the activities that are done by the database administrator and users during the pre-language operational stage. The operation of the database query system is depicted in the system architecture, as shown in Figure 4. The pre-language operational stage will be conducted using a normal computing terminal over the Internet, while the query operations will be handled through query formulation using the mobile phones and query processing by the application and database servers.

EXPERIMENTAL STUDY

The prototype of the query language is developed to identify the query interface midlet on the mobile phones and J2SE for the query engine servlets on the application server. The database consists of data on students, subjects, and staff of the university. Students enroll in many subjects that are conducted in many sessions at several venues. The subjects are taught by many lecturers who are of different ranks. The students stay in hostels managed by wardens, and their outings must be approved by a staff member. The schema in each relation in the database is shown in Figure 5 and their relationships in Table 1 respectively.

The database is used to prove that the developed query language is relationally complete. As mentioned earlier, there are five fundamental operations that must be satisfied in order for a language to be considered relationally complete.

Table 1. The relationships between attributes

ATTRIBUTE	REFER TO
Hostelwarden	Staff ID
Staff position	StaffPost code
Staff ranking	StaffRank code
Student hostel Code	Hostel code
Subject lecturer	Staff ID
Session subjectCode	Subject code
Session venue	Venue code
Outing studentID	Student ID
Outing approver	Staff ID
Enrolment studentID	Student ID
Enrolment subjectCode	Subject code

In order to explain the operations, let us assume that there are two relations *R* and *S* which contain m and n numbers of attributes respectively. The five operations are: *Selection, Projection, Cartesian Product, Union,* and *Set Difference.* Here we describe how the five operations are handled by the developed query language:

- **Selection:** Mathematically denoted as $\sigma_{predicate}(R)$, works on a single relation *R* and defines a relation that contains only those tuples of *R* that satisfy the specified condition (predicate). For example, a *Selection* query of $\sigma_{studYear=3}(Student)$ will produce as output, tuples from the *Student* relation which have a value 3 in their *studYear* attribute.
  The developed query language handles the *Selection* operation by allowing a user to select the proper relation name from the list of keywords. This action will allow the user to later choose the name of the attribute that he/she wants to check the value of, to choose the operation he/she wants to perform, and to provide the value he/she is looking for. Figure 6 show the screen shots of a sample *Selection* operation. The query language also allows multiple conditions to be implemented.
- **Projection:** Mathematically denoted as $\pi_{a1,a2,\ldots,am}(R)$, works on a single relation *R* and defines a relation that contains a vertical subset of *R*, extracting the values of specified attributes and eliminating duplicates. For example, a *Projection* query of $\pi_{subjCode,subjNam\,e}(Subject)$ will produce as output all tuples in the *Subject* relation. For each tuple, the only values associated with the attributes of *subjCode* and *subjName* will be returned. The developed query language handles the *Projection* operation by allowing a user to select the proper relation name which later gives as a list all of the attributes of the relation. A user can then select as many attributes as he/she likes to view as output. Figure 7 shows the screen shots of a sample *Projection* operation.
- **Cartesian Product:** Mathematically denoted as *R X S*, defines a relation that is the concatenation of every tuple of relation *R* with every tuple of relation *S*. For example, if a *Staff* relation contains 100 tuples with 10 attributes each and a *Subject* relation contains 100 tuples with five attributes each, a *Cartesian product* query of *Staff X Subject* will produce as output 1,000 times 100 tuples, since each tuple of the *Staff* relation will be concatenated with each one of the tuples from the *Subject* relation. Furthermore, each tuple of the output relation will have ten plus five attributes. The result of a *Cartesian product* operation is less meaningful. Therefore, a more restrictive form of the operation, called *Join*, is more preferable. A *Join* operation, mathematically denoted as $R \triangleright\triangleleft_{predicate} S$, includes only the combinations of both relations that satisfy certain conditions. For example, a query of *Staff* $\triangleright\triangleleft_{staffID=lectID}$ *Subject* will produce tuples which combine a tuple from the *Staff* with its associated *Subject* tuple. The number of the output tuples will be equal to the number of the tuples of *Staff.* There are several variations to the *Join* operation such as *left-outer join,*

Figure 5. The database schema

Hostel (code, name, warden)
Staff (ID, name, gender, DOB, mobile#, position, ranking, area)
StaffPost (code, name)
StaffRank (code, name)
Student (ID, name, gender, DOB, year, hostelCode, room)
Subject (code, name, creditHour, lecturer)
Session (subjectCode, day, timeStart, timeEnd, venue)
Venue (code, name, capacity)
Enrolment (studentID, subjectCode)
Outing (studentID, dateOut, tiemout, destination, dateIn, timeIn, approver)

Figure 6. Selection operation

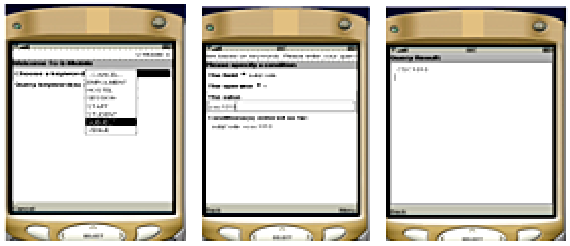


Figure 7.

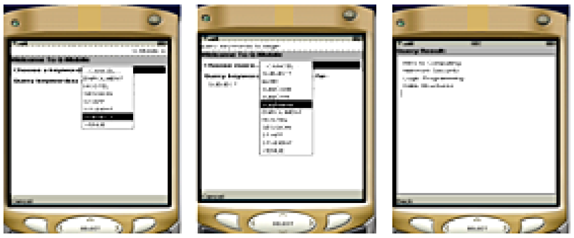
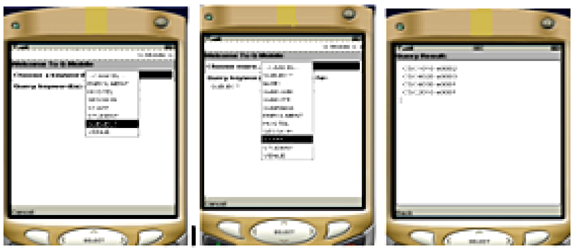


Figure 8.



right-outer join, equijoin, and so on. The developed query language handles the *Join* operation by allowing a user to select as many relations as he/she wants to join and the relations can be indirectly related as well. Figure 8 shows screen shots of a sample *Join* operation.

- **Union:** Mathematically denoted as $R \cup S$, concatenates all tuples of R and all tuples of S into one relation with duplicate tuples being eliminated. However, R and S must be union-compatible (i.e., both relations contain the same number of attributes, and each corresponding attribute is of the same domain) in order for the operation to be valid. For example, the query of *Lecturing-Staff* $\cup$ *Administrative-Employee* is valid if they both have the same schema type. This operation is not yet implemented by the query language. But the concept would be possibly implemented by allowing a user to select two relations and the union operator. The query engine will then check for the compatibility of their schema types.
- **Set Difference:** Mathematically denoted as $R - S$, defines a relation consisting of the tuples that are in relation R, but not in S. R and S again must be union-compatible. For example, the query of *Staff–Lecturing-Staff* will produce all administrative staff, assuming lecturers and administrators are the only two types of staff in the university. This operation is not yet implemented by the query language. Similarly, the concept would be possibly implemented by allowing a user to select two relations and a difference operator. The query engine will then check for the compatibility of their schema types. Besides the five operations, the query language is capable of combining multiple operations in one query, and it can also be easily extended to include other operations as needed.

CONCLUSION

The use of a keyword-based query language with menu-based guidance for formulating database queries using mobile phones is feasible due to its usability, powerfulness, flexibility, and scalability. With the physical constraints of the mobile phones, this type of query language uses minimal space for presentation and a lesser number of interactions to form complex queries. Furthermore, the keyword-based language is robust since it enables users to enter all possible queries by combining relevant keywords. Therefore, it is able to accept unplanned queries; it can be extended, and it is adaptable to other database applications. With further research, especially in the method for recommending the keywords relevant to a user, the keyword-based language could be the answer to access a full-scale database from mobile phones.

REFERENCES

Agrawal, S., Chaudhuri, S., & Das, G. (2002, February 26-March 1). DBXplorer: A system for keyword-based search over relational databases. *Proceedings of the IEEE 18th International Conference on Data Engineering (ICDE'02)* (pp. 5-16).

Bai, B. R., Chen, C. L., Chien, L. F., & Lee, L. S. (1998). Intelligent retrieval of dynamic networked information from mobile terminals using spoken natural language queries. *IEEE Transactions on Consumer Electronics, 44*(1), 62-72.

Boonsrimuang, P., Kobayashi, H., & Paungma, T. (2002, July 3-5). Mobile Internet navigation system. *Proceedings of the 5th IEEE International Conference on High Speed Networks and Multimedia Communications* (pp. 325-328).

Calado, P., da Silva, A.S., Laender, A. H. F., Ribeiro-Neto, B. A., & Vieira, R. C. (2004). A Bayesian Network approach to searching Web databases through keyword-based queries. *Information Processing and Management, 40*(5), 773-790.

Cao, G. (2002). On improving the performance of cache invalidation in mobile environments. *Mobile Networks and Applications, 7*(4), 291-303.

Catarci, T. (1991). On the expressive power of graphical query languages. *Proceedings of the 2nd IFIP W.G. 2.6 Working Conference on Visual Databases* (pp. 404-414).

Chandra, A. (1988). Theory of database queries. *Proceedings of the 7th ACM Symposium on Principles of Database Systems* (pp. 1-9).

Chang, E., Seide, F., Meng, H. M., Chen, Z., Shi, Y., & Li, Y. C. (2002). A system for spoken query information retrieval on mobile devices. *IEEE Transactions on Speech and Audio Processing, 10*(8), 531-541.

Connolly, T., Begg, C., & Strachan, A. (1997). *Database systems—A practical approach to design, implementation and management.* Boston: Addison-Wesley.

Czejdo, B., Rusinkiewicz, M., Embley, D., & Reddy, V. (1989, October 4-6). A visual query language for an ER data model. *Proceedings of the IEEE Workshop on Visual Languages* (pp.165-170).

Hongchi, S., Shang, Y., & Ren, F. (2001, October 7-10). Using natural language to access databases on the Web. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 1, pp. 429-434).

Hung, K., & Zhang, Y.-T. (2003). Implementation of a WAP-based telemedicine system for patient monitoring. *IEEE Transactions on Information Technology in Biomedicine, 7*(2), 101-107.

Kang, I.-S., Bae, J.-H. J., & Lee, J. H. (2002, November 6-8). Database semantics representation for natural language access. *Proceedings of the 1st International Symposium on Cyber Worlds* (pp. 127-133).

Kimoto, H., & Iwadera, T. (1991, July 8-14). A dynamic thesaurus and its application to associated information retrieval. *Proceedings of IJCNN-91, the Seattle International Joint Conference on Neural Networks* (Vol. 1, pp. 19-29).

Koyama, A., Takayama, N., Barolli, L., Cheng, Z., & Kamibayashi, N. (2002, November 6-8). An agent based campus information providing system for cellular phone. *Proceedings of the 1st International Symposium on Cyber Worlds* (pp. 339-345).

Massari, A., Weissman, S., & Chrysanthis, P. K. (1996). Supporting mobile database access through query by icons. *Distributed and Parallel Databases (Special Issue on Databases and Mobile Computing), 4*(3), 47-68.

Myers, B. A., & Beigl, M. (2003). Handheld computing. *Computer, 36*(9), 27-29.

Polyviou, S., Samaras, G., & Evripidou, P. (2005). A relationally complete visual query language for heterogeneous data sources and pervasive querying. *Proceedings of IEEE 2005.*

Ramakrishnan, R., & Gehrke, J. (2000). *Database management systems.* New York: McGraw-Hill.

Tan, K. L., & Ooi, B. C. (1998). Batch scheduling for demand-driven servers in wireless environments. *Journal of Information Sciences, 109*(1-4), 281-298.

Weibenberg, N., Voisard, A., & Gartmann, R. (2006). Using ontologies in personalized mobile applications. *Proceedings of the12th Annual International Workshop on GIS.* ACM Press.

Knowledge Representation in Semantic Mobile Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

Mobile applications today face the challenges of increasing information, diversity of users and user contexts, and ever-increasing variations in mobile computing platforms. They need to continue being a successful business model for service providers and useful to their user community in the light of these challenges.

An appropriate representation of information is crucial for the agility, sustainability, and maintainability of the information architecture of mobile applications. This article discusses the potential of the Semantic Web (Hendler, Lassila, & Berners-Lee, 2001) framework to that regard.

The organization of the article is as follows. We first outline the background necessary for the discussion that follows and state our position. This is followed by the introduction of a knowledge representation framework for integrating Semantic Web and mobile applications, and we deal with both social prospects and technical concerns. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

In recent years, there has been a proliferation of affordable information devices such as a cellular phone, a personal digital assistant (PDA), or a pager that provide access to mobile applications. In a similar timeframe, the Semantic Web has recently emerged as an extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

The goal of the mobile Web is to be able to mimic the desktop Web as closely as possible, and an appropriate representation of information is central to its realization. This requires a transition from the traditional approach of merely presentation to *representation* of information. The Semantic Webprovides one avenue towards that.

Indeed, the integration of Semantic Webtechnologies in mobile applications is suggested in Alesso and Smith (2002) and Lassila (2005). There are also proof-of-concept semantic mobile applications such as MyCampus (Gandon & Sadeh, 2004) and mSpace Mobile (Wilson, Russell, Smith, Owens, & Schraefel, 2005) serving a specific community. However, these initiatives are limited by one or more of the following factors: the discussion of knowledge representation is one-sided and focuses on specific technology(ies) or is not systematic, or the treatment is restricted to specific use cases. One of the purposes of this article is to address this gap.

Table 1. Knowledge representation tiers in a semantic mobile application

Semiotic Level	Semantic Mobile Web Concern and Technology Tier	Decision Support
Social	Trust	Feasibility
Pragmatic	Inferences	
Semantic	Metadata, Ontology, Rules	
Syntactic	Markup	
Empirical	Characters, Addressing, Transport	
Physical	Not Directly Applicable	

UNDERSTANDING KNOWLEDGE REPRESENTATION IN SEMANTIC MOBILE APPLICATIONS

In this section, our discussion of semantic mobile applications is based on the knowledge representation framework given in Table 1.

The first column addresses semiotic levels. Semiotics (Stamper, 1992) is concerned with the use of symbols to convey knowledge. From a semiotics perspective, a representation can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social, each depending on the previous one in that order. The physical level is concerned with the representation of signs in hardware and is not directly relevant here.

The second column corresponds to the Semantic Web"tower" that consists of a stack of technologies that vary across the technical to social spectrum as we move from bottom to top, respectively. The definition of each layer in this technology stack depends upon the layers beneath it.

Finally, in the third column, we acknowledge that there are time, effort, and budgetary constraints on producing a

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

representation and include feasibility as an all-encompassing factor on the layers to make the framework practical. For example, an organization may choose not to adopt a technically superior technology as it cannot afford training or processing tools available that meet the organization's quality expectations. For that, analytical hierarchy process (AHP) and quality function deployment (QFD) are commonly used techniques. Further discussion of this aspect is beyond the scope of the article.

The architecture of a semantic mobile application extends that of a traditional mobile application on the server-side by: (a) expressing information in a manner that focuses on *description* rather than presentation or processing of information, and (b) associating with it a knowledge management system (KMS) consisting of one or more domain-specific ontologies and a reasoner.

We now turn our attention to each of the levels in our framework for knowledge representation in semantic mobile applications.

Empirical Level of a Semantic Mobile Application

This layer is responsible for the communication properties of signs. Among the given choices, the Unicode Standard provides a suitable basis for the signs themselves and is character-by-character equivalent to the ISO/IEC 10646 Standard Universal Character Set (UCS). Unicode is based on a large set of characters that are needed for supporting internationalization and special symbols. This is necessary for the aim of universality of mobile applications.

The characters must be uniquely identifiable and locatable, and thus addressable. The uniform resource identifier (URI), or its successor international resource identifier (IRI), serves that purpose.

Finally, we need a transport protocol such as the hypertext transfer protocol (HTTP) or the simple object access protocol (SOAP) to transmit data across networks. We note that these are limited to the transport between the mobile service provider that acts as the intermediary between the mobile client and the server. They are also layered on top of and/or used in conjunction with other protocols, such as those belonging to the Institute of Electrical and Electronics Engineers (IEEE) 802 hierarchy.

Syntactic Level of a Semantic Mobile Application

This layer is responsible for the formal or structural relations between signs. The eXtensible Markup Language (XML) lends a suitable syntactical basis for expressing information in a mobile application.

The XML is supported by a number of ancillary technologies that strengthen its capabilities. Among those, there are domain-specific XML-based markup languages that can be used for expressing information in a mobile application (Kamthan, 2001).

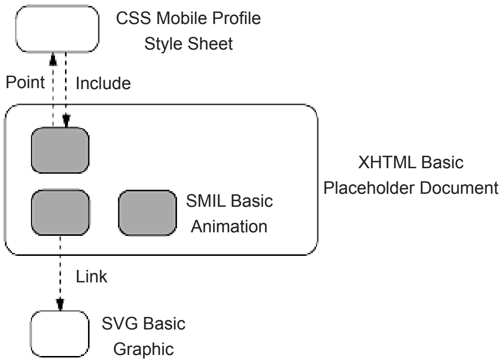
The eXtensible HyperText Markup Language (XHTML) is a recast of the HyperText Markup Language (HTML) in XML. XHTML Basic is the successor of compact HTML (cHTML) that is an initiative of the NTT DoCoMo, and of the Wireless Markup Language (WML) that is part of the wireless application protocol (WAP) architecture and an initiative of the Open Mobile Alliance (OMA). It uses XML for its syntax and HTML for its semantics. XHTML Basic has native support for elementary constructs for structuring information like paragraphs, lists, and so on. It could also be used as a placeholder for information fragments based on other languages, a role that makes it rather powerful in spite of being a small language.

The Scalable Vector Graphics (SVG) is a language for two-dimensional vector graphics that works across platforms, across output resolutions, across color spaces, and across a range of available bandwidths; SVG Tiny and SVG Basic are profiles of SVG targeted towards cellular phones and PDAs, respectively.

The Synchronized Multimedia Integration Language (SMIL) is a language that allows description of temporal behavior of a multimedia presentation, associates hyperlinks with media objects, and describes the layout of the presentation on a screen. It includes reusable components that can allow integration of timing and synchronization into XHTML and into SVG. SMIL Basic is a profile that meets the needs of resource-constrained devices such as mobile phones and portable disc players.

Namespaces in XML is a mechanism for uniquely identifying XML elements and attributes of a markup language, thus making it possible to create heterogeneous (compound) documents (Figure 1) that unambiguously mix

Figure 1. The architecture of a heterogeneous XML document for a mobile device



elements and attributes from multiple different XML document fragments.

Appropriate presentation on the user agent of information in a given modality is crucial. However, XML in itself (and by reference, the markup languages based on it) does not provide any special presentation semantics (such as fonts, horizontal and vertical layout, pagination, and so on) to the documents that make use of it. This is because the separation of the structure of a document from its presentation is a design principle that supports maintainability of a mobile application. The cascading style sheets (CSS) provides the presentation semantics on the client, and CSS mobile profile is a subset of CSS tailored to the needs and constraints of mobile devices.

With the myriad of proliferating platforms, information created for one platform needs to be adapted for other platforms. The eXtensible Stylesheet Language Transformations (XSLT) is a style sheet language for transforming XML documents into other, including non-XML, documents. As an example, information represented in XML could be transformed on-demand using an XSLT style sheet into XHTML Basic or an SVG Tiny document, as appropriate, for presentation to users accessing a mobile portal via a mobile device.

Representing information in XML provides various advantages towards archival, retrieval, and processing. It is possible to down-transform and render a document on multiple devices via an XSLT transformation, without making substantial modifications to the original source document. However, XML is not suitable for completely representing the knowledge inherent in information resources. For example, XML by itself does not provide any specific mechanism for differentiating between homonyms or synonyms, does not have the capabilities to model complex relationships precisely, is not able to extract implicit knowledge (such as hidden dependencies), and can only provide limited reasoning and inference capabilities, if at all.

The combination of the layers until now forms the basis of the mobile Web. The next two layers extend that and are largely responsible for what could be termed as the semantic mobile Web.

Semantic Level of a Semantic Mobile Application

This layer is responsible for the relationship of signs to what they stand for. The resource description framework (RDF) is a language for metadata that provides a "bridge" between the syntactic and semantic layers. It, along with RDF Schema, provides elementary support for *classification* of information into classes, properties of classes, and means to model more complex relationships among classes than possible with XML only. In spite of their usefulness, RDF/RDF Schema suffer from limited representational capabilities and non-standard semantics. This motivates the need for additional expressivity of knowledge.

The declarative knowledge of a domain is often modeled using ontology, an explicit formal specification of a conceptualization that consists of a set of concepts in a domain and relations among them (Gruber, 1993). By explicitly defining the relationships and constraints among the concepts in the universe of discourse, the *semantics* of a concept is constrained by restricting the number of possible interpretations of the concept.

In recent years, a number of initiatives for ontology specification languages for the semantic Web, with varying degrees of formality and target user communities, have been proposed, and the Web Ontology Language (OWL) has emerged as the successor. Specifically, we advocate that OWL DL, one of the sub-languages of OWL, is the most suitable among the currently available choices for representation of domain knowledge in mobile applications due to its compatibility with the architecture of the Web in general; and the Semantic Webin particular benefits from using XML/RDF/RDF Schema as its serialization syntax, its agreement with the Web standards for accessibility and internationalization, well-understood declarative semantics from its origins in description logics (DL) (Baader, McGuinness, Nardi, & Schneider, 2003), and provides the necessary balance between computational expressiveness and decidability.

Pragmatic Level of a Semantic Mobile Application

This layer is responsible for the relation of signs to interpreters. There are several advantages of an ontological representation. When information is expressed in a form that is oriented towards presentation, the traditional search engines usually return results based simply on a string match. This can be ameliorated in an ontological representation where the search is based on a *concept* match. An ontology also allows the logical means to distinguish between homonyms and synonyms, which could be exploited by a reasoner conforming to the language in which it is represented. For example, Java in the context of coffee is different from that in the context of an island, which in turn is different from the context of a programming language; therefore a search for one should not return results for other. Further, ontologies can be applied towards precise access of desirable information from mobile applications (Tsounis, Anagnostopoulos, & Hadjiefthymiades, 2004). Even though resources can be related to one another via a linking mechanism, such as the XML Linking Language (XLink), these links are merely structural constructs based on author discretion that do not carry any special semantics.

Explicit declaration of all knowledge is at times not cost effective as it increases the size of the knowledge base, which

Example 1. Ontological Inferences

```
<Region rdf:ID=“MontTremblant”>
 <subRegionOf rdf:resource=“#Laurentides”/>
</Region>
<Region rdf:ID=“Laurentides”>
 <subRegionOf rdf:resource=“#Qu&eacute;bec”/>
</Region>
<owl:TransitiveProperty rdf:ID=“subRegionOf”>
 <rdfs:domain rdf:resource=“#Region”/>
 <rdfs:range rdf:resource=“#Region”/>
</owl:TransitiveProperty>
```

becomes rather challenging as the amount of information grows. However, an ontology with a suitable semantical basis can make implicit knowledge (such as hidden dependencies) *explicit.* A unique aspect of ontological representation based for instance on OWL DL is that it allows logical constraints that can be reasoned with and enables us to *derive* logical consequences—that is, facts not literally present in the ontology but *entailed* by the semantics.

We have a semantic mobile portal for tourist information. Let Mont Tremblant, Laurentides, and Québec be defined as regions, and the subRegionOf property between regions be declared as transitive in OWL (see Example 1.)

Then, an OWL reasoner should be able to derive that if Mont Tremblant is a sub-region of Laurentides, and Laurentides is a sub-region of Québec, then Mont Tremblant is also a sub-region of Québec. This would give a more complete set of search results to a semantic mobile application user.

In spite of its potential, ontological representation of information presents certain domain-specific and human-centric challenges (Kamthan & Pai, 2006). Query formulations to a reasoner for extracting information from an ontology can be rather lengthy input on a mobile device. It is currently also difficult both to provide a sound logical basis to aesthetical, spatial/temporal, or uncertainty in knowledge, and represent that adequately in ontology.

Social Level of a Semantic Mobile Application

This layer is responsible for the manifestation of social interaction with respect to the representation. Specifically, ontological representations are a result of consensus, which in turn is built upon trust.

The client-side environment in a mobile context is constrained in many ways: devices often have restricted processing capability and limited user interface input/output facilities. The Composite Capabilities/Preference Profiles (CC/PP) Specification, layered on top of XML and RDF, allows the expression of user (computing environment and personal) preferences, thereby informing the server side of the delivery context.

Example 2. Device Profile

```
<rdf:RDF xmlns:rdf=“http://www.w3.org/1999/02/22-rdf-syntax-ns#”
 xmlns:ccpp=“http://www.w3.org/2002/11/08-ccpp-schema#”
 xmlns:prf=“http://a.com/schema#”>
 . . .
 <ccpp:component>
 <rdf:Description rdf:about=“http://a.com/HardwareDevice”>
 <rdf:type rdf:resource=“http://a.com/schema#HardwarePlatform”/>
 <ccpp:defaults rdf:resource=“http://a.com/HardwareDefault”/>
 <prf:vendor>MyMobileCompany</prf:vendor>
 <prf:cpu>ABC</prf:cpu>
 <prf:displayHeight>200</prf:displayHeight>
 <prf:displayWidth>320</prf:displayWidth>
 <prf:memoryMb>16</prf:memoryMb>
 </rdf:Description>
 </ccpp:component>
 . . .
</rdf:RDF>
```

In Example 2, CC/PP markup for a device whose processor is of type ABC and the preferred default values of its display and memory as determined by its vendor are given. The namespace in XML is used to disambiguate elements/attributes that are native to CC/PP or RDF from those that are specific to the vendor vocabulary.

CC/PP can be used as a basis for introducing context-awareness in mobile applications (Sadeh, Chan, Van, Kwon, & Takizawa, 2003; Khushraj & Lassila, 2004).

One of the major challenges to the personalization based on profile mechanism is the user concern for privacy. The Platform for Privacy Preferences Project (P3P) allows the expression of privacy preferences of a user, which can be used by agents to decide if they have the permission to process certain content, and if so, how they should go about it. This ensures that users are informed about privacy policies of the mobile service providers before they release personal information. Thus, P3P provides a balance to the flexibility offered by the user profiles in CC/PP.

The Security Assertion Markup Language (SAML), XML Signature, and XML Encryption provide assurance of the sanctity of the message to processing agents. We note that an increasing number of languages to account for may place an unacceptable load, if it is to be processed exclusively, on the client side. We also acknowledge that these technologies alone will not solve the issue of trust, but when applied properly, could contribute towards it.

FUTURE TRENDS

The transition of the traditional mobile applications to semantic mobile applications is an important issue. The previous section has shown the amount of expertise and level of skills required for that. Although up-transformations are in general difficult, we anticipate that the move will be easier for the mobile applications that are well-structured in their current expression of information and in their conformance to the languages deployed.

The production of mobile applications, and by extension semantic mobile applications, is becoming increasingly complex and resource (time, effort) intensive. Therefore, a systematic and disciplined approach for their development, deployment, and maintenance, similar to that of Web engineering, is needed. Related to that, the issue of quality of represented and delivered information will continue to be important. The studies of specific attributes such as usability (Bertini, Catarci, Kimani, & Dix, 2005) and "best practices" for mobile applications from the World Wide Web Consortium (W3C) Mobile Web Initiative are efforts that could eventually be useful in an "engineering" approach for producing future semantic mobile applications.

The process of aggregation and inclusion of information in a mobile application is primarily manual, which can be both tedious and error prone. This process could be, at least partially, automated via the use of Web services where mobile applications could be made to automatically update themselves with (candidate) information. Therefore, manifestations of mobile applications through Semantic Webservices (Wagner & Paolucci, 2005; Wahlster, 2005) are part of a natural evolution.

CONCLUSION

For mobile applications to continue to provide a high quality-of-service (QoS) to their user community, their information architecture must be evolvable. The incorporation of Semantic Webtechnologies can be much more helpful in that regard. The adoption of these technologies does not have to be an "all or nothing" proposition: the evolution of a mobile application to a semantic mobile application could be gradual, transcending from one layer to another in the aforementioned framework. In the long term, the benefits of transition outweigh the costs.

Ontologies form one of the most important layers in semantic mobile applications, and ontological representations have certain distinct advantages over other means of representing knowledge. However, engineering an ontology is a resource-intensive process, and an ontology is only as useful as the inferences (conclusions) that can be drawn from it.

To be successful, semantic mobile applications must align themselves to the Semantic Webvision of inclusiveness for all. For that, the semiotic quality of representations, particularly that of ontologies, must be systematically assured and evaluated.

REFERENCES

Alesso, H. P., & Smith, C. F. (2002). *The intelligent wireless Web.* Boston: Addison-Wesley.

Baader, F., McGuinness, D., Nardi, D., & Schneider, P. P. (2003). *The description logic handbook: Theory, implementation and applications.* Cambridge University Press.

Bertini, E., Catarci, T., Kimani, S., & Dix, A. (2005). A review of standard usability principles in the context of mobile computing. *Studies in Communication Sciences, 1*(5), 111-126.

Gandon, F. L., & Sadeh, N. M. (2004, June 1-3). Context-awareness, privacy and mobile access: A Web semantic and multiagent approach. *Proceedings of the 1st French-Speaking Conference on Mobility and Ubiquity Computing* (pp. 123-130), Nice, France.

Gruber, T.R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *Formal ontology in conceptual analysis and knowledge representation.* Kluwer Academic.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The semantic Web. *Scientific American, 284*(5), 34-43.

Kamthan, P. (2001, March 20-22). Markup languages and mobile commerce: Towards business data omnipresence. *Proceedings of the WEB@TEK 2001 Conference,* Québec City, Canada.

Kamthan, P., & Pai, H.-I. (2006, May 21-24). Human-centric challenges in ontology engineering for the Semantic Web: A perspective from patterns ontology. *Proceedings of the 17th Annual Information Resources Management Association International Conference (IRMA 2006)*, Washington, DC.

Khushraj, D., & Lassila, O. (2004, November 7). CALI: Context Awareness via Logical Inference. *Proceedings of the Workshop on Semantic Web Technology for Mobile and Ubiquitous Applications,* Hiroshima, Japan.

Lassila, O. (2005, August 25-27). Using the Semantic Web in ubiquitous and mobile computing. *Proceedings of the 1st International IFIP/WG 12.5 Working Conference on Industrial Applications of the Semantic Web (IASW 2005)*, Jyväskylä, Finland.

Sadeh, N. M., Chan, T.-C., Van, L., Kwon, O., & Takizawa, K. (2003, June 9-12). A Semantic Web environment for context-aware m-commerce. *Proceedings of the 4th ACM Conference on Electronic Commerce* (pp. 268-269), San Diego, CA.

Stamper, R. (1992, October 5-8). Signs, organizations, norms and information systems. *Proceedings of the 3rd Australian*

Conference on Information Systems (pp. 21-55), Wollongong, Australia.

Tsounis, A., Anagnostopoulos, C., & Hadjiefthymiades, S. (2004, September 13). The role of Semantic Web and ontologies in pervasive computing environments. *Proceedings of the Workshop on Mobile and Ubiquitous Information Access (MUIA 2004)*, Glasgow, Scotland.

Wagner, M., & Paolucci, M. (2005, June 9-10). Enabling personal mobile applications through Semantic Web services. *Proceedings of the W3C Workshop on Frameworks for Semantics in Web Services*, Innsbruck, Austria.

Wahlster, W. (2005, June 3). Mobile interfaces to intelligent information services: Two converging megatrends. *Proceedings of the MINDS Symposium,* Berlin, Germany.

Wilson, M., Russell, A., Smith, D. A., Owens, A., & Schraefel, M. C. (2005, November 7). mSpace mobile: A mobile application for the Semantic Web. *Proceedings of the 2nd International Workshop on Interaction Design and the Semantic Web,* Galway, Ireland.

KEY TERMS

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Knowledge Representation: The study of how knowledge about the world can be represented and the kinds of reasoning can be carried out with that knowledge.

Ontology: An explicit formal specification of a conceptualization that consists of a set of terms in a domain and relations among them.

Personalization: A strategy that enables delivery that is customized to the user and user's environment.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and their representations.

User Profile: An information container describing user needs, goals, and preferences.

Location-Based Multimedia Content Delivery System for Monitoring Purposes

Athanasios-Dimitrios Sotiriou
National Technical University of Athens, Greece

Panagiotis Kalliaras
National Technical University of Athens, Greece

INTRODUCTION

Advances in mobile communications enable the development and support of real-time multimedia services and applications. These can be mainly characterized by the personalization of the service content and its dependency to the actual location within the operational environment. Implementation of such services does not only call for increased communication efficiency and processing power, but also requires the deployment of more intelligent decision methodologies.

While legacy systems are based on stationary cameras and operational centers, advanced monitoring systems should be capable of operating in highly mobile, ad-hoc configurations, where overall situation and users roles can rapidly change both in time and space, exploiting the advances in both the wireless network infrastructure and the user terminals' capabilities. However, as the information load is increased, an important aspect is its filtering. Thus, the development of an efficient rapid decision system, which will be flexible enough to control the information flow according to the rapidly changing environmental conditions and criteria, is required. Furthermore, such a system should interface and utilize the underlying network infrastructures for providing the desired quality of service (QoS) in an efficient manner.

In this framework, this article presents a *location-based multimedia content delivery system* (LMCDS) for monitoring purposes, which incorporates media processing with a decision support system and positioning techniques for providing the appropriate content to the most suitable users, in respect to user profile and location, for monitoring purposes. This system is based on agent technology (Hagen & Magendanz, 1998) and aims to promote the social welfare, by increasing the overall situation awareness and efficiency in emergency cases and in areas of high importance. Such a system can be exploited in many operational (public or commercial) environments and offers increased security at a low cost.

SERVICES

The LMCDS provides a platform for rapid and easy set up of a monitoring system in any environment, without any network configurations or time-consuming structural planning. The cameras can be installed in an ad hoc way, and video can be transmitted to and from heterogeneous devices using an Intelligent decision support system (IDSS) according to the user's profile data, location information, and network capabilities.

Users can dynamically install ad-hoc cameras to areas where the fixed camera network does not provide adequate information. The real-time transmission of still images or video in an emergency situation or accident to the available operational centers can instantly provide the necessary elements for the immediate evaluation of the situation and the deployment of the appropriate emergency forces. This allows the structure of the monitoring system to dynamically change according to on-the-spot needs.

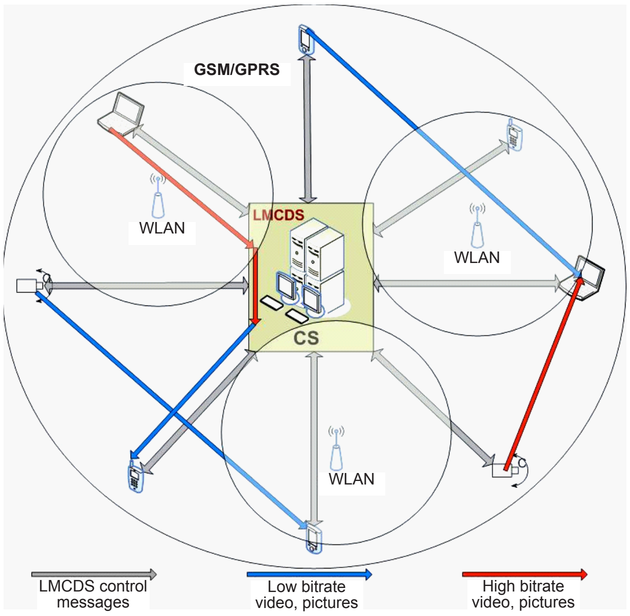
The IDSS is responsible for overviewing the system's activity and providing multimedia content to the appropriate users. Its functionality lies in the following actions:

- identifying the appropriate user or group of users that need access to the multimedia content (either through user profile criteria or topological criteria);
- providing the appropriate multimedia content in relevance to the situation and the location; and
- adapting the content to the user's needs due to the heterogeneity of the users devices—that is, low bit rate video to users with portable devices.

The LMCDS can evaluate users' needs and crisis events in respect to the topological taxonomy of all users and provide multimedia content along with geographical data. The location information is obtained through GPS or from GPRS through the use of corresponding techniques (Markoulidakis, Desiniotis, & Kypris, 2004). It also provides intelligent methodologies for processing the video and image content according to network congestion status and terminal devices. It can handle the necessary monitoring management mechanisms, which enable the selection of the non-congested network elements for transferring the appropriate services (i.e., video streaming, images, etc.) to the concerned users. It also delivers the service content in the most appropriate

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. System functionality and services



format, thus allowing the cooperation of users equipped with different types of terminal devices.

Moreover, the LMCDS provides notification services between users of the system for instant communication in case of emergency through text messaging or live video feed.

All of the above services outline the requirements for an advanced monitoring system. The LMCDS functionality meets these requirements, since it performs the following features:

- location-based management of the multimedia content in order to serve the appropriate users;
- differentiated multimedia content that can be transmitted to a wide range of devices and over different networks;
- lightweight codecs and decoders that can be supported by devices of different processing and network capabilities;
- IP-based services in order to be transparent to the underlying network technology and utilize already available hardware and operating systems platforms;
- intelligent delivery of the multimedia content through the LMCDS in order to avoid increased traffic payload as well as information overload; and
- diverse localization capabilities through both GPS and GPRS, and generation of appropriate topological data (i.e., maps) in order to aid users.

However, the system architecture enables the incorporation of additional location techniques (such as WLAN positioning mechanisms)) through the appropriate, but simple, development of the necessary interfaces with external location mechanisms.

In order to describe the above services in a more practical way, a short list of available options and capabilities of the system is given below. The target group of the LMCDS consists of small to medium businesses that seek a low-cost solution for monitoring systems or bigger corporations that need an ad hoc extension to their current system for emergency cases and in order to enhance their system's mobility. Even though the users of the system consist mainly of security staff and trained personnel that are in charge of security, the system's ease of use, low user complexity, and device diversity allow access even to common untrained users.

The system offers a range of capabilities, most of which are summarized in Figure 1, such as:

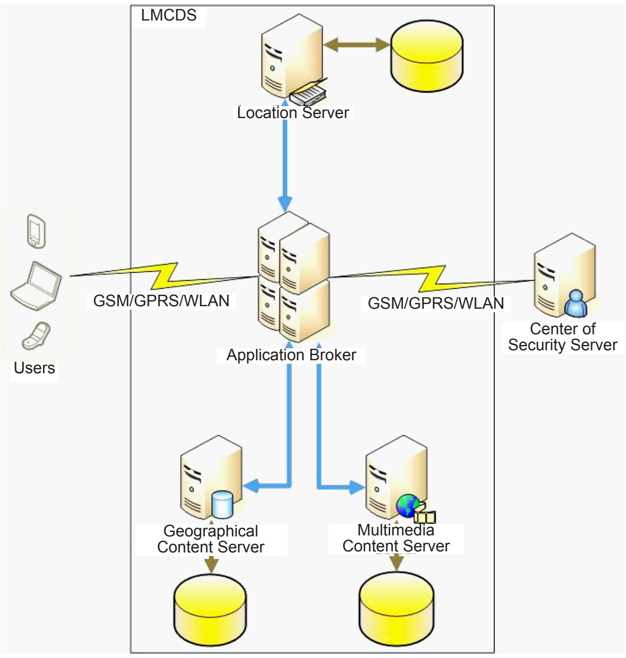
- User registration and authentication.
- User profile (i.e., device, network interface).
- Location awareness:
    - User is located through positioning techniques.
    - User is presented with appropriate topographical information and metadata.
    - User is aware of all other users' locations.
    - User can be informed and directed from a Center of Security (CS) to specified locations.
- Multimedia content:
    - Video, images, and text are transmitted to user in real time or off-line based on situation or topological criteria.
    - User can provide feedback from his device through camera (laptop, PDA, smart phone) or via text input.
    - Content is distributed among users from the CS as needed.
- Ad hoc installation of cameras that transmit video to CS and can take advantage of wireless technology (no fixed network needed).
- Autonomous nature of users due to agent technology used.

LMCDS ARCHITECTURE

The LMCDS is designed to distribute system functionality and to allow diverse components to work independently while a mass amount of information is exchanged. This design ensures that new users and services can be added in an ad hoc manner, ensuring future enhancements and allowing it to support existing monitoring systems.

Multi-agent systems (MASs) (Nwana & Ndumu, 1999) provide an ideal mechanism for implementing such a heterogeneous and sophisticated distributed system in contrast to traditional software technologies' limitations in communication and autonomy.

Figure 2. Architecture overview


LMCDS
Location Server
Users
GSM/GPRS/WLAN
Application Broker
GSM/GPRS/WLAN
Center of Security Server
Geographical Content Server
Multimedia Content Server

The system is developed based on MAS and allows diversified agents to communicate with each other in an autonomous manner, resulting in an open, decentralized platform. Tasks are being delegated among different components based on specific rules, which relate to the role undertaken by each agent, and information is being composed to messages and exchanged using FIPA ACL (FIPA, 2002a). An important aspect for the communication model is the definition of the content language (FIPA, 2002b). Since LMCDS targets to a variety of devices, including lightweight terminals, the LEAP language (Berger, Rusitschka, Toropov, Watzke, & Schlichte, 2002) of the JADE technology has been exploited.

In addition to the security mechanisms supported by the underlying network components, the JADE platform offers a security model (Poggi, Rimassa, & Tomaiuolo, 2001) that enables the delegation of tasks to respective agent components by defining certificates, and ensures the authentication and encryption of TCP connections through the secure socket layer (SSL) protocol (http://www.openssl.org/).

The general architecture of the LMCDS is shown in Figure 2. The platform is composed of different agents offering services to the system which are linked by an application broker agent, acting as the coordinator of the system. These agents are the location server, the center of security server, the application broker, the geographical content server, and the multimedia content server. The latter two are discussed in later sections in more detail, while a brief description of the functionality of the others is given as follows.

The location server agent is responsible for the tracking of all users and the forwarding of location-based information to other components. The information is gathered dynamically and kept up-to-date according to specific intervals. The intelligence lies in the finding of the closest users to the demanded area, not only in terms of geographical coordinates, but also in terms of the topology of the environment. More information on location determination is given in a following section.

The center of security server agent monitors all users and directs information and multimedia content to the appropriate users. It is responsible for notifying users in emergency situations, and also performs monitoring functions for the system and its underlying network.

The User agent components manage information, including the transmission and reception of image or video, the display of location information, critical data, or other kinds of displays. They are in charge of several user tasks, such as updating users' preferences and location, decoding the response messages from the application broker, and performing service level agreement (SLA) (ADAMANT, 2003) monitoring functionalities. The user agent can reside in a range of devices, since it is Java based.

Finally, the application broker agent acts as a mediator that provides the interface between all of the components. It is responsible for prioritizing user requests according to users' profiles, performing authentication functions, and acting as a facilitator for SLA negotiations. It also coordinates the communication process between users in order for the multimedia content to be delivered in the appropriate format according to the user's processing and network capabilities, and also the network's payload.

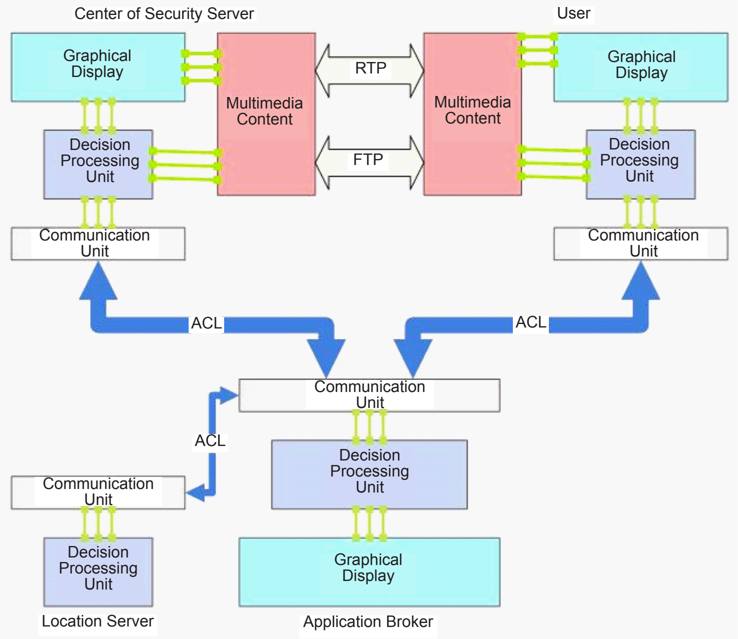
A closer look into the components of each agent, along with the interaction between them, is shown in Figure 3. Each agent is composed of three main components: the graphical display, which is responsible for user input and information display; the communication unit, in charge of agent communication through ACL messages; and the decision processing unit, which processes all received information. In addition, the user and the CS include a multimedia content component for the capture, playback, and transmission of multimedia data through RTP or FTP channels.

MULTIMEDIA PROCESSING AND ENCODING

Video/Image Formats

One of the novelties is the ability of the system to perform real-time format conversions of the image or video data and transmit to several heterogeneous recipients, ranging from large PC servers to small personal devices like PDAs or mobile phones.

Figure 3. Agent components


Center of Security Server
User
Graphical Display
Multimedia Content
RTP
FTP
Multimedia Content
Graphical Display
Decision Processing Unit
Decision Processing Unit
Communication Unit
Communication Unit
ACL
ACL
Communication Unit
ACL
Communication Unit
Decision Processing Unit
Decision Processing Unit
Graphical Display
Location Server
Application Broker

The LMCDS enables the adoption and support of different video formats, according to the partial requirements of the user terminals and the available networks status. The most commonly used is the M-JPEG (http://www.jpeg .org/index.html*)* format. It was preferred over other common video formats suitable for wireless networks like MPEG (http://www. m4if.org/*)* and H.263 (http://www.itu.int/rec/recommendation.asp), which provide higher compression ratio, because using them can require intense processing power both for the encoder and the decoder. Also, frames in MPEG or H.263 streams are inter-related, so a single packet loss during transmission may degrade video quality noticeably. On the contrary, M-JPEG is independent of such cascaded-like phenomena, and it is preferable for photo-video application temporal compression requirements for smoothness of motion.

It is a lossy codec, but the compression level can be set at any desired quality, so the image degradation can be minimal. Also, at small data rates (5-20Kbps) and small frame rates, M-JPEG produces better results than MPEG or H.263. This is important, as the photos or video can be used as clues in legal procedures afterwards, where image quality is more crucial than smooth motion. Another offering feature is the easy extraction of JPEG (http://www.jpeg .org/index.html*)* images from video frames.

The video resolution can be set in any industry-standard (i.e., subQCIF) or any other resolution of width and height dividable of 8, so the track is suitable for the device it is intended for. Video is streamed directly from the camera-equipped terminals in a peer-to-peer manner. Transmission rates for the video depend on the resolution and the frame rate used. Some sample rates are given in Table 1.

Apart from M-JPEG, another set of video formats have been adopted, such as H263 and MPEG-4. The development of these formats enables the testing and evaluation of the LMCDS, based on the network congestion and the current efficiency of the supported video formats and the crisis situations in progress. This means that for a specific application scenario, the encoding with M-JPEG format can lead to better quality on the user terminal side, while the MPEG 4 format can be effective in cases that the network infrastructure is highly loaded, so the variance in bit rate can keep the quality in high values.

Image compression is JPEG with resolution of any width and height dividable of 8. For the real-time transmission of video stream, the real-time transfer protocol (RTP, http://www.ietf.org/rfc/rfc3550.txt) is used, while for stored images and video tracks, the file transfer protocol (FTP) is used.

Video Processing

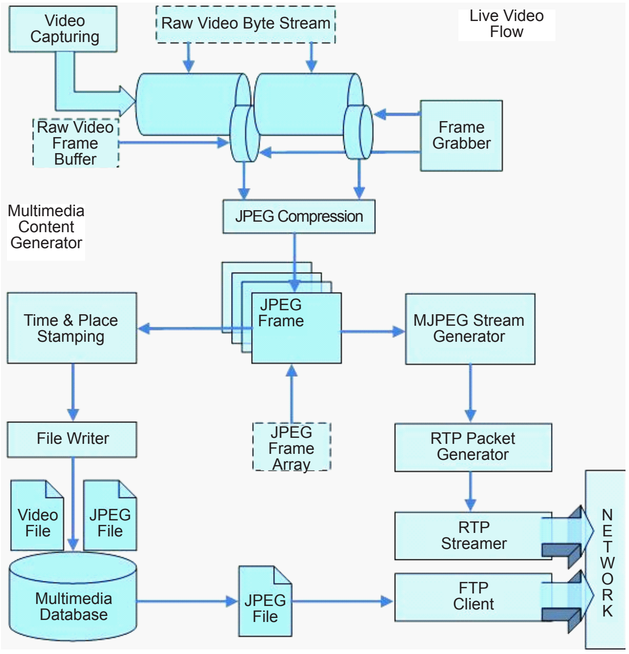
It is important to point out that the output video formats can be produced and transmitted simultaneously with the use of the algorithm shown in Figure 4.

Note that the image/video generator can be called several times for the same captured video stream as long as it is fed with video frames from the frame grabber. So, a single user can generate multiple live video streams with variations,

Table 1. Output video formats for the application

Resolution	Frame Rate	Suitable Network	Trans. Rate (kbps)	Target Device
160 x 120	1	GPRS WLAN	2 – 3	Smartphone PDA, PC
160 x 120	5	GPRS WLAN	10 – 15	Smartphone, PDA,PC
232 x 176	2	GPRS WLAN	10 – 15	PDA,PC
320 x 240	5	WLAN	30 – 40	PC
320 x 240	15	WLAN	90 – 120	PC
640 x 480	5	WLAN	45 – 55	PC

Figure 4. LMCDS video processing overview



not only in frame rate and size, but also in JPEG compression quality, color depth, and even superimpose layers with handmade drawings or text. The algorithm was implemented in Java with the use of the JMF API (http://java.sun. com/products/java-media/jmf/).

For a captured stream at a frame rate of *n* frames per second, the frame grabber component extracts from the raw video byte stream *n*/*A* samples per second, where *A* is a constant representing the processing power of the capturing device. Depending on how many different qualities of video streams need to be generated, *m* Image/Video Generator processes are activated, and each process *i* handles K_i fps. The following relationship needs to be applied:

$$K_i = B_i * n / A * m \text{ , where}$$

$$B_i = (K_i * A * m)/ n \text{ and}$$

$$\sum B_i = m$$

There is also a latency of $m*C_i$ seconds per video stream, where C_i depends on the transcoding time of the image to each final format.

So, this tendency of the LMCDS to keep the frame rate low is inevitable due to this sampling process. However, using low frame rates is quite common in surveillance systems. It also allows long hours of recording, where video size is optimal when minimum both for storage and transmission, and is less demanding in processing power for use with video players running on small devices.

GEOGRAPHICAL CONTENT DELIVERY

Positioning Methods

The LMCDS uses the following techniques for locating the users' positions inside the served environment.

- *GPS (Global Positioning System)* is a global satellite system based on a group of non-geostatic satellites in middle altitude orbit (12,000 miles). The GPS-enabled devices have the ability to locate their position with a high degree of accuracy by receiving signals from at least six satellites.
- *A GSM/GPRS subscriber can be located upon request, depending on the received power from adjacent cells. More information about these mechanisms can be found in* Markoulidakis et al. (2004).
- The *Ekahau* position engine (http://www.ekahau.com/pdf/EPE_2.1_datasheet.PDF) is designed to locate 802.11b (at present) wireless LAN users positions in the indoor environment. In the context of radio resource management, the software could apply the access point's radio coverage map in the indoor environment. It uses a calibration method. Initially it measures a set of sample points' radio strength. Based on these sample points, the engine can estimate a client WLAN station's approximate location, given an arbitrary combination of signal strengths.

Geographical Database

The geographical database of the LMCDS is storing information about the positions of the users that are registered in the system. The information is obtained regularly by scheduled queries. When the users perform service request messages to the system, their position coordinates are automatically retrieved (through any of the above methods), so their location in the geographical DB is also updated. Special importance has been given to the ability of the system to serve queries about the relative position of its users and estimation of distances. So, the scheduled queries can be of the following types:

- Find my 3 nearest users and send them a video.
- Estimate the time that I need to get to Building A.
- *Find the users closest to point B and send pictures to those that operate in GPRS network and high-quality video to those in UMTS or WLAN.*
- *When user C enters a specified area, send him a message.*

The user is displayed visual information in the form of maps to its device. The map consists of raster data—that is, the plan of the area and also several layers of metadata, showing points of interest, paths, as well as the position of relative users.

FUTURE TRENDS

Future steps involve the exploitation of video streaming measurements for providing guaranteed QoS of the video content to the end user, as well as the better utilization of the available radio resources. Furthermore, the incorporation of new trends in video streaming in conjunction with a markup language for multimedia content, such as MPEG-7 or MPEG-21, can offer a higher level of personalized location-based services to the end user and are in consideration for future development.

CONCLUSION

This article presented a location-based multimedia content system enabling real-time transfer of multimedia content to end users for location-based services. Based on the general architecture of multi-agent systems, it focused on fundamental features that enable the personalization of the service content and the intelligent selection of the appropriate users for delivering the selected content.

REFERENCES

Berger, M., Rusitschka, S., Toropov, D., Watzke, M., & Schlichte, M. (2002). Porting distributed agent-middleware to small mobile devices. *Proceedings of the Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices,* Bologna, Italy.

FIPA (Foundation for Intelligent Physical Agents). (2002a). *FIPA ACL message structure specification.* SC00061G.

FIPA. (Foundation for Intelligent Physical Agents). (2002b). *FIPA SL content language specification.* SC00008I.

Hagen, L., & Magendanz, T. (1998). Impact of mobile agent technology on mobile communication system evolution. *IEEE Personal Communications, 5*(4).

IST ADAMANT Project. (2003). *SLA management specification.* IST-2001-39117, Deliverable D6.

Markoulidakis, J. G., Desiniotis, C., & Kypris, K. (2004). Statistical approach for improving the accuracy of the CGI++ mobile location technique. *Proceedings of the Mobile Location Workshop , Mobile Venue '04.*

Nwana, H., & Ndumu, D. (1998). A perspective on software agents research. *The Knowledge Engineering Review, 14*(2).

Poggi, A., Rimassa, G., & Tomaiuolo, M. (2001). Multi-user and security support for multi-agent systems. *Proceedings of WOA 2001 Workshop,* Modena, Italy.

KEY TERMS

Agent: A program that performs some information gathering or processing task in the background. Typically, an agent is given a very small and well-defined task.

Application Broker: A central component that helps build asynchronous, loosely coupled applications in which independent components work together to accomplish a task. Its main purpose is to forward service requests to the appropriate components.

IDSS: Intelligent decision support system.

LMCDS: Location-based multimedia content delivery system.

MAS: Multi-agent system.

Media Processing: Digital manipulation of a multimedia stream in order to change its core characteristics, such as quality, size, format, and so forth.

Positioning Method: One of several methods and techniques for locating the exact or relative geographical position of an entity, such as a person or a device.

Location-Based Multimedia Services for Tourists

Panagiotis Kalliaras
National Technical University of Athens, Greece

Athanasios-Dimitrios Sotiriou
National Technical University of Athens, Greece

P. Papageorgiou
National Technical University of Athens, Greece

S. Zoi
National Technical University of Athens, Greece

INTRODUCTION

The evolution of mobile technologies and their convergence with the Internet enable the development of interactive services targeting users with heterogeneous devices and network infrastructures (Wang et al., 2004). Specifically, as far as cultural heritage and tourism are concerned, several systems offering location-based multimedia services through mobile computing and multimodal interaction have already appeared in the European research community (LOVEUS, n.d.; Karigiannis, Vlahakis, & Daehne, n.d.).

Although such services introduce new business opportunities for both the mobile market and the tourism sector, they are not still widely deployed, as several research issues have not been resolved yet, and also available technologies and tools are not mature enough to meet end user requirements. Furthermore, user heterogeneity stemming both from different device and network technologies is another open issue, as different versions of the multimedia content are often required.

This article presents the AVATON system. AVATON aims at providing citizens with ubiquitous user-friendly services, offering personalized, location-aware (GSM Association, 2003), tourism-oriented multimedia information related to the area of the Aegean Volcanic Arc. Towards this end, a uniform architecture is adopted in order to dynamically release the geographic and multimedia content to the end users through enhanced application and network interfaces, targeting different device technologies (mobile phones, PDAs, PCs, and TV sets). Advanced positioning techniques are applied for those mobile user terminals that support them.

SERVICES

AVATON is an ambient information system that offers an interactive tour to the user (visitor) in the area of the Aegean Volcanic Arch (see http://www.aegean.gr/petrified_forest/). The system can serve both as a remote and as an onsite assistant for the visitor, by providing multimedia-rich content through various devices and channels:

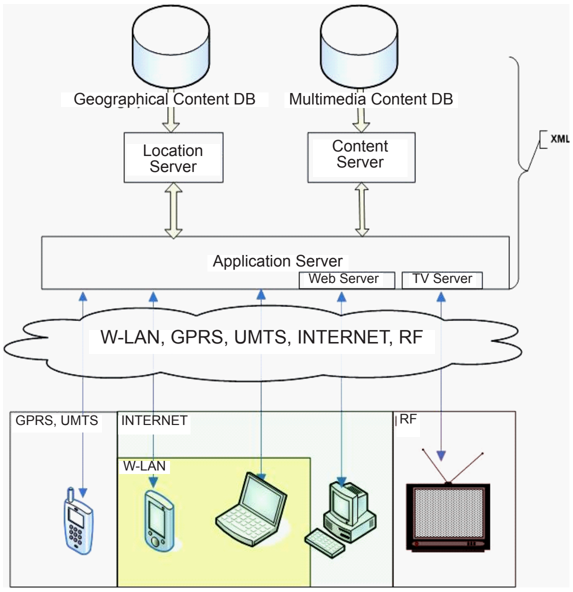
- over the Internet, via Web browsers with the use of new technologies such as rich-clients and multi-tier architecture in order to dynamically provide the content;
- with portable devices (palmtops, PDAs) and 2.5G or 3G mobile phones, which are capable of processing and presenting real-time information relevant to the user's physical position or areas of interest; and
- via television channels—AVATON allows users to directly correlate geographic with informative space and conceivably pass from one space to the other, in the context of Worldboard (Spohrer, 1999).

With the use of portable devices equipped with positioning capabilities, the system provides:

- dynamic search for geographical content, information related to users' location, or objects of interest that are in their proximity;
- tours in areas of interest with the aid of interactive maps and 3-D representations of the embossed geography;
- search for hypermedia information relative to various geographic objects of the map;
- user registration and management of personal notes during the tour that can be recalled and reused during later times; and

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. The AVATON architecture


Geographical Content DB
Multimedia Content DB
Location Server
Content Server
XML
Application Server
Web Server
TV Server
W-LAN, GPRS, UMTS, INTERNET, RF
GPRS, UMTS
INTERNET
W-LAN
RF

- interrelation of personal information with natural areas or objects for personal use or even as a collective memory relative to visited areas or objects.

THE AVATON ARCHITECTURE

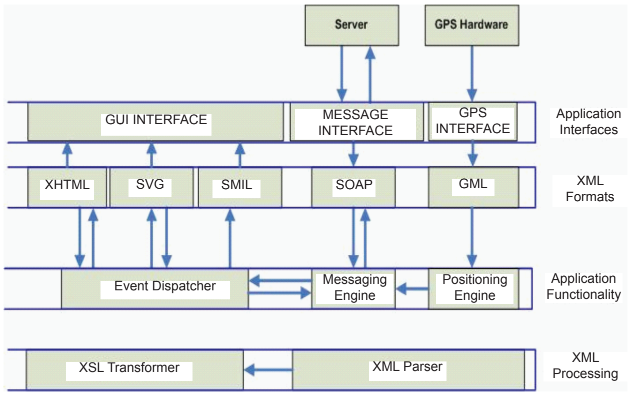
Overview

The AVATON system is based on a client-server architecture composed of three main server components: the application server, the content server, and the location server. The application server combines information and content from the content and location servers, and replies to client requests through different network technologies. The content is retrieved from two kinds of databases, the geographical and multimedia content DBs. The above architecture is shown in Figure 1.

In more detail:

- **Multimedia Content Database:** This database contains the multimedia content such as images, video, audio, and animation.
- **Geographical Content Database:** A repository of geographical content such as aerial photos, high-resolution maps, and relevant metadata.
- **Content Server:** The content server supplies the application server with multimedia content. It retrieves needed data from the multimedia content database according to user criteria and device capabilities, and responds to the application server.

Figure 2. The XML-based technologies in the client side


Server
GPS Hardware
GUI INTERFACE
MESSAGE INTERFACE
GPS INTERFACE
Application Interfaces
XHTML
SVG
SMIL
SOAP
GML
XML Formats
Event Dispatcher
Messaging Engine
Positioning Engine
Application Functionality
XSL Transformer
XML Parser
XML Processing

- **Location Server:** Serves requests for geographical content from the application server by querying the geographical content database. The content retrieved is transformed into the appropriate format according to user device display capabilities and network bandwidth available.
- **Application Server:** The application server receives requests from different devices through GPRS, UMTS (third-generation mobile phone), W-LAN, Internet (PDA, laptop, PC), and RF (television). The server identifies each device and transmits data in an appropriate format. More precisely, the application server incorporates a Web server and a TV server in order to communicate with PCs and televisions respectively.

Client

This section focuses on the mobile-phone and PDA applications. The scope of the AVATON system includes Java-enabled phones with color displays and PDAs with WLAN or GPRS/UMTS connectivity. While all the available data for the application can be downloaded and streamed over the network, data caching is exploited for better performance and more modest network usage.

When the users complete their registration in the system, they have in their disposal an interactive map that initially portrays the entire region as well as areas or individual points of interest. For acquiring user position, the system is using GPS. The client also supports multi-lingual implementation, as far as operational content is concerned, for example menus, messages, and help. These files are maintained as XML documents. XML is extensively used in order to ease the load of parsing different data syntaxes. A single process, the XML Parser is used for decoding all kinds of data and an XSL Transformer for transcoding them in new formats. The different XML formats are XHTML, SVG, SMIL, SOAP, and GML, as shown in Figure 2.

Geographical Info Presentation

In order to render the geographical data, the client receives raster images for the drawing of the background map, combined with metadata concerning areas of interest and links to additional textual or multimedia information. The raster data are aerial high-resolution photographs of the region on two or three scales. Because of the high resolution of the original images, the client is receiving small portions, in the form of tiles from the *raster data processing engine* in the server side, which are used to regenerate the photorealistic *image layer* in a resolution that is suitable for the device used. The attributes of the geographical data are generated in vector ShapeFile (ESRI ShapeFile) format, which is quite satisfactory for the server side but not for lightweight client devices. So, a *SHP TO SVG converter* at the server side is regenerating the metadata in SVG format that can be viewed properly from a handheld device. As soon as the metadata is downloaded to the client device, a final filtering (XSLT transformation) is done and the additional layer is opposed to the image layer in the *SVG viewer.* On the *SVG data layer,* the user can interact with points of interest and receive additional information in the form of text or multimedia objects. The above are shown in Figure 3.

Multimedia Info Presentation

The presentation of multimedia information mainly depends on user position. The system is designed to provide audio and video clips, 3-D representations, and also textual information concerning each place of interest. Not all devices, though, receive the same content, since they differ in display, processor, or network speed. For that purpose, for each registered user, the system decides what kind of content is more suitable for them to receive and the multimedia content server generates the appropriate script. Depending on the available memory of the client's device, media objects stay resident in the cache memory so that frequently requested content is accessed without delays that occur due to network latency. In Figure 4 the components that are involved in the multimedia presentation are shown. The *TourScript Data* contains the script which describes the multimedia presentation. It is transcoded inside the *SMIL generator* to a SMIL message that follows the XML syntax, so that it can be incorporated seamlessly to the messages that are exchanged in the AVATON system. At the client side, the SMIL message is received by the *SMIL processor* which coordinates the process of fetching the *multimedia objects* from the *client cache memory* to the suitable renderer, so that the *multimedia presentation* can be completed.

Location Server

The location server is the component that handles the geographical content of the AVATON system. It provides a storage system for all geographical data and allows querying of its contents through location criteria, such as global position and areas of interest. Content management is based on a PostgreSQL (http://www.postgres.org) relational database. A JDBCInterface uses the JDBC APIs in order to provide support for data operations. A GISExtension is also present, based on PostGIS (http://www.postgis.org), in order to enable the PostgreSQL server to allow spatial queries. This feature is utilized through a GISJDBCInterface, a PostGIS layer on top of PostgreSQL.

Figure 3. Client-side map rendering

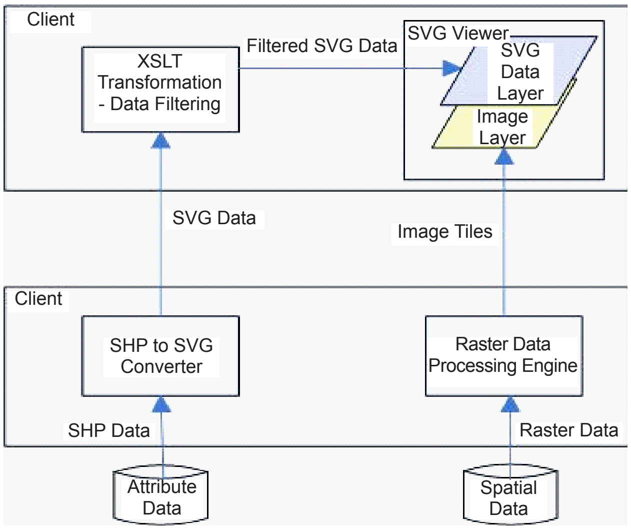
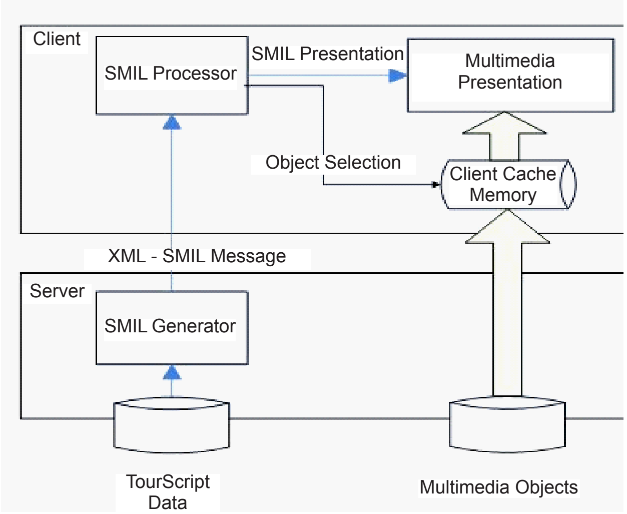

Client
XSLT Transformation - Data Filtering
Filtered SVG Data
SVG Viewer
SVG Data Layer
Image Layer
SVG Data
Image Tiles
Client
SHP to SVG Converter
Raster Data Processing Engine
SHP Data
Raster Data
Attribute Data
Spatial Data

Figure 4. Client-side map rendering


Client
SMIL Processor
SMIL Presentation
Multimedia Presentation
Object Selection
Client Cache Memory
XML - SMIL Message
Server
SMIL Generator
TourScript Data
Multimedia Objects

Cartographic Data

Concerning the photorealistic information, the user can choose from several distinct zoom levels. The mobiles phones and PDAs in the market that support GPRS or WLAN have displays of different resolutions that, in most cases, are multiples of 16 pixels. Hence, the location server can generate tiles with a multiple of 16p x 16p, which can be presented in the user's mobile device. The server always holds multiple resolutions for every level of cartographic (photographic) information. The levels of cartographic information define the degree of focus.

Apart from the photographic layer, additional layers of vector information also exist, and their size is approximately 5% or 10% of the corresponding photographic. Therefore, in practice, every device will initially request from the server the cartographical information with the maximum resolution this device can support. Hence, the server decides the available resolution that best corresponds to the requested resolution from the device. The size of every tile is approximately 1K. The devices with greater resolution per tile receive more files, with greater magnitude, for every degree of focus due to the higher resolution.

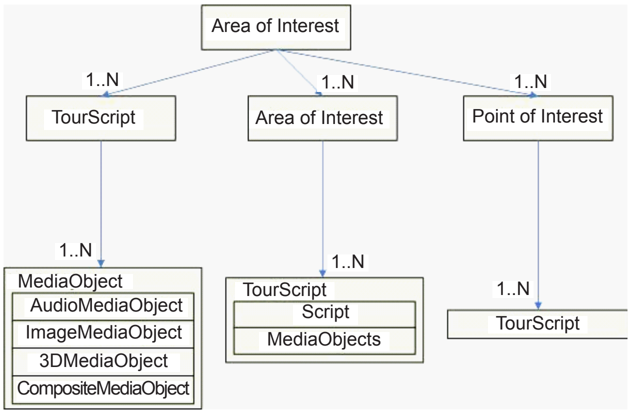
Multimedia Content Server

The multimedia content server component comprises the major unit that controls the mixing and presentation of different multimedia objects. Its purpose is to upload all the objects necessary and present them in a well-defined controlled order that in general depends on the user position, interactions, and tracking information available. The multilingual audiovisual information scheduled for presentation is coordinated so that several objects may be presented simultaneously. The multimedia content server component is also responsible for choosing "relevant" objects for the user to select among in the case the user requires more information on a topic. The multimedia content server interfaces with the multimedia content database, a relational database storing the multimedia content. The database is organized thematically and allows the creation of hierarchical structures. It also contains a complete list of multimedia material, covering all content of the physical site, such as 3D reconstructed plants, audio narration, virtual 3D models, avatar animations, and 2D images.

Media Objects

As mentioned already, the multimedia content server is responsible for mixing the basic units of multimedia information. These elements are hierarchically ordered. At the finest level of granularity, there are atomic objects called MediaObjects with specializations such as AudioMediaObject, ImageMediaObject, 3DMediaObject, and CompositeMediaObject. These objects contain the actual data to be rendered along with additional profile metadata characterizing them. At a higher level of complexity, a TourScript represents an ordered sequence of MediaObjects, all of which are to be presented if the script is chosen.

Figure 5. Hierarchy formulation of media objects at the multimedia content server



According to user requirements*,* the user will be able to navigate through the site in a geographically based tree. This is made possible through the use of points of interest (PoI) and areas of interest (AoI). A PoI can only contain TourScripts and can be viewed as the end node of the site tree. In contrast, an AoI may contain either another PoI, an AoI, or TourScripts. This allows the system to map the actual site into a hierarchy model containing PoI at the top and MediaObject components at the leaf level.

The multimedia content server is also responsible for managing this site-tree for the entire site. Moreover it is responsible for traversing it. The use of the site tree is quite interesting: when a media object, for instance *audio* object, is presented to the user, it belongs to a node in the site hierarchy. Figure 5 shows the structure of the site in a tree view as described previously. The multimedia content server is responsible for coordinating the rendering components in order to provide a synchronized presentation to the user, according to user preferences, position, and commands.

Deployment and Usage

Based on the proposed architecture, the AVATON services are being deployed to physical sites within the Aegean Volcano Arc (such as Santorini and Lesvos islands) and evaluated by real end users under different scenarios. The main air interfaces that will be used by the system (along

Figure 6. Implementation plan for the Lesvos island site

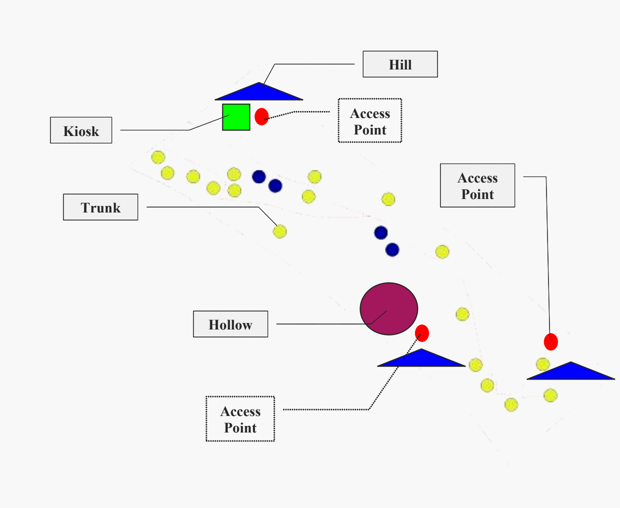
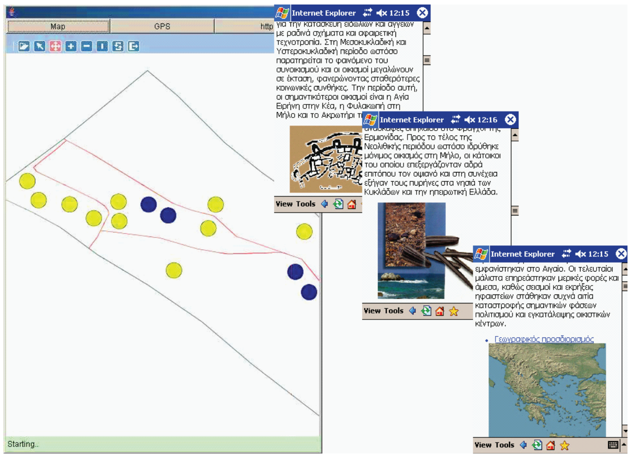


Figure 7. SVG map and interface



για την κατασκευή ειδωλίων και αγγείων με ραδινά σχήματα και αφαιρετική τεχνοτροπία. Στη Μεσοκυκλαδική και Υστεροκυκλαδική περίοδο ωστόσο παρατηρείται το φαινόμενο του συνοικισμού και οι οικισμοί μεγαλώνουν σε έκταση, φανερώνοντας σταθερότερες κοινωνικές συνθήκες. Την περίοδο αυτή, οι σημαντικότεροι οικισμοί είναι η Αγία Ειρήνη στην Κέα, η Φυλακωπή στη Μήλο και το Ακρωτήρι τ

Ερμιονίδας. Προς το τέλος της Νεολιθικής περιόδου ωστόσο ιδρύθηκε μόνιμος οικισμός στη Μήλο, οι κάτοικοι του οποίου επεξεργάζονταν αδρά επιτόπου τον οψιανό και στη συνέχεια εξήγαν τους πυρήνες στα νησιά των Κυκλάδων και την ηπειρωτική Ελλάδα.

εμφανίστηκαν στο Αιγαίο. Οι τελευταίοι μάλιστα επηρεάστηκαν μερικές φορές και άμεσα, καθώς σεισμοί και εκρήξεις ηφαιστείων στάθηκαν συχνά αιτία καταστροφής σημαντικών φάσεων πολιτισμού και εγκατάλειψης οικιστικών κέντρων.

- Γεωγραφικός προσδιορισμός

with standard wired access through common LANs or the Internet) are:

- **WLAN:** A standard 802.11b wireless access network provides connectivity for users equipped with portable devices (such as PDA with built-in WLAN cards or laptops).
- **GPRS/UMTS:** For mobile users and smart phones, access will be provided through the GSM network, using GPRS (Bettsteter, Vögel, & Eberspächer, 1999). This restricts the system from providing video or 3-D animations to such users, and the services offered are focused on text, images (including map information), and short audio. As GPRS is already packet oriented, our implementation can be easily transferred to UMTS, if available.

In Figure 6, the actual implementation plan is given for the location of the Sigri Natural History Museum on Lesvos island. The area consists of an open geological site, the Petrified Forest, where the ash from a volcanic eruption some 15 to 20 million years ago covered the stand of sequoia trees, causing their petrification. Wireless access is provided by the use of a 3 Netgear 54Mbps access point equipped with additional Netgear antennas in order to overcome the physical limitations of the area (hills, trunks, and hollows).

Visitors to the site are equipped with PDAs or smart phones (provided at the entrance kiosk) and stroll around the area. A typical scenario consists of the following: The users enter the archaeological site and activate their devices. They then perform a login and provide personal details to the server, such as username, language selection, and device settings. The client then requests from the server and loads the map of the area in SVG (http://www.w3.org/TR/SVG) format, as seen in Figure 7. The circles on the map present distinct points of interest (yellow indicating trees and blue indicating leafs). The application monitors the users' location and updates the SVG map in real time, informing the users of their position. The SVG map is interactive, and when the users enter the vicinity of a point of interest, the application automatically fetches and displays (via their browser or media player) the corresponding multimedia content (in the form of HTML pages, audio, or video) at the requested language. The users are also able to navigate manually through the available content and receive additional information on topics of their interest. During the tour, the users' path is being tracked and displayed (the red line on Figure 7) in order to guide them through the site. They are also able to keep notes or mark favorite content (such as images), which can be later sent to them when they complete the tour.

FURTHER WORK

The system is currently being deployed and tested in two archeological sites. Users are expected to provide useful feedback on system capabilities and assist in further enhancements of its functionality. Also, as 3G infrastructure is being expanded, incorporation of the UMTS network in the system's access mechanisms will provide further capabilities for smart phone devices and also use of the system in areas where wireless access cannot be provided.

Towards commercial exploitation, billing and accounting functionalities will be incorporated into the proposed architecture. Finally, possible extensions of the system are considered in order to include other cultural or archeological areas.

REFERENCES

Bettsteter, C., Vögel, H.-J., & Eberspächer, J. (1999). GSM Phase 2+ General Packet Radio Service GPRS: Architecture, protocols, and air interface. *IEEE Communication Surveys,* (3rd Quarter).

ESRI ShapeFile Technical Description. (1998, July). *An ESRI white paper.*

GSM Association. (2003, January). Location based services. *SE.23, 3.10.*

Karigiannis, J. N., Vlahakis, V., & Daehne, P. (n.d.). ARCHEOGUIDE: Challenges and solutions of a personalized augmented reality guide for archeological sites. *Computer Graphics in Art, History and Archeology, Special Issue of the IEEE Computer Graphics and Application Magazine.*

LOVEUS. (n.d.). Retrieved from http://loveus.intranet.gr/documentation.htm

Spohrer, J. C. (1999). Information in places. *IBM System Journal, 38*(4).

Wang, Y., Cuthbert, L., Mullany, F. J., Stathopoulos, P., Tountopoulos, V., Sotiriou, D. A., Mitrou, N., & Senis, M. (2004). Exploring agent-based wireless business models and decision support applications in an airport environment. *Journal of Telecommunications and Information Technology,* (3).

KEY TERMS

Cartographic Data: Spatial data and associated attributes used by a geographic information system (GIS).

Content Provider: A service that provides multimedia content.

Location-Based Services: A way to send custom advertising and other information to cell phone subscribers based on their curent location.

Multimedia: Media that usesmultiple forms of information content and information processing (e.g., text, audio, graphics, animation, video, interactivity) to inform or entertain the (user) audience.

Network: A network of telecommunications links arranged so that data may be passed from one part of the network to another over multiple links.

Tourism: The act of travel for predominantly recreational or leisure purposes, and the provision of services in support of this act.

Location-Based Services

Péter Hegedüs
Budapest University of Technology and Economics, Hungary

Mihály Orosz
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The basically two different technologies, the location-based services in the mobile communication and the well-elaborated multicast technology, are joined in the multicast over LBS solutions. As the article demonstrates, this emerging and new management area has many possibilities that have not been completely utilized.

Currently an important area of mobile communications is the *ad-hoc computer networking,* where mobile devices need base stations; however, they form an overlay without any Internet-related infrastructure, which is a virtual computer network among them. In their case the selective, location-related communication model has not been elaborated on completely (Ibach, Tamm, & Horbank, 2005). One of the various communication ways among the software entities on various mobile computers is the one-to-many data dissemination that is called *multicast.* Multicast communication over mobile ad-hoc networks has increasing importance. This article describes the fundamental concepts and solutions, especially focusing on the area of *location-based services* (LBSs) and the possible multicasting over the LBS systems. This kind of communication is in fact a special case of the multicast communication model, called *geocast,* where the sender disseminates the data to that subset of the multicast group members in a specific geographical area. The article shows that the geocast utilizes the advantages of the LBS, since it is based on the location-aware information being available in the location-based solutions (Mohapatra, Gui, & Li, 2004).

There are several unsolved problems in LBS, in management and low surfaces. Most of them are in quick progress, but some need new developments. The *product managers* have to take responsibility for the software and hardware research and development part of the LBS product. This is a very important part of the design process, because if the development engineer leaves the product useful out of consideration, the whole project could possibly be led astray. Another import question is that of LBS-related *international* and *national laws,* which could throw an obstacle into LBS's spread. These obstacles will need to be solved before LBS global introduction.

The article presents this emerging new area and the many possible management solutions that have not been completely utilized.

BACKGROUND

Location-based services are based on the various distances of mobile communications from different base stations. With advances in automatic position sensing and wireless connectivity, the application range of mobile LBS is rapidly developing, particularly in the area of geographic, tourist, and local travel information systems (Ibach et al., 2005). Such systems can offer maps and other area-related information. The LBS solutions offer the capability to deliver location-aware content to subscribers on the basis of the positioning capability of the wireless infrastructure. The LBS solutions can push location-dependent data to mobile users according to their interests, or the user can pull the required information by sending a request to a server that provides location-dependent information.

LBS may have many useful applications in homeland security (HLS). A few of the more significant of these applications are security and intelligence operations, notification systems for emergency responders, search and rescue, public notification systems, and emergency preparedness (Niedzwiadek, 2002). Mobile security and intelligence operatives can employ LBS to aid in monitoring people and resources in space and time, and they can stay connected with emergency operations centers to receive the necessary updates regarding the common operating picture for a situation. Emergency operations centers can similarly coordinate

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

search and rescue operations. Call-down systems can be employed to notify the public in affected disaster areas. In this application the multicast communication is preferable, since it uses in an efficient way the communication channels, which can be partly damaged after a disaster. Location-based public information services can give time-sensitive details concerning nearest available shelters, safe evacuation routes, nearest health services, and other public safety information (Niedzwiadek, 2002).

Location-based services utilize their ability of location-awareness to simplify user interactions. With advances in wireless connectivity, the application range of mobile LBSs is rapidly developing, particularly in the field of tourist information systems—telematic, geographic, and logistic information systems. However, current LBS solutions are incompatible with each other since manufacturer-specific protocols and interfaces are applied to aggregate the various system components for positioning, networking, or payment services. In many cases, these components form a rigid system. If such system has to be adapted to another technology, for example, moving from *global positioning system* (GPS)-based positioning to in-house *IEEE 802.11a*-based *wireless local-area network* (WLAN) or *Bluetooth*-based positioning, it has to be completely redesigned (Haartsen, 1998). In such a way the ability of interoperation of different resources under changeable interconnection conditions becomes crucial for the end-to-end availability of the services in mobile environments (Ibach, & Horbank, 2004).

There are a lot of location determination methods and technologies, such as the satellite-based GPS, which is widely applied (Hofmann-Wellenhof, Lichtenegger, & Collins, 1997). The three basic location determination methods are *proximity, triangulation* (lateration), and *scene analysis* or *pattern recognition* (Hightower & Borriello, 2001). Signal strength is frequently applied to determine proximity. As a proximity measurement, if a signal is received at several known locations, it is possible to intersect the coverage areas of that signal to calculate a location area. If one knows the angle of bearing (relative to a sphere) and distance from a known point to the target device, then the target location can be accurately calculated. Similarly, if somebody knows the range from three known positions to a target, then the location of the target object can be determined. A GPS receiver uses range measurements to multiple satellites to calculate its position. The location determination methods can be *server based* or *client based* according to the locus of computation (Hightower & Borriello, 2001).

Chen et al. (2004) introduce an enabling infrastructure, which is a middleware in order to support location-based services. This solution is based on a *location operating reference model* (LORE) that solves many problems of constructing location-based services, including location modeling, positioning, tracking, location-dependent query processing, and smart location-based message notification. Another interesting solution is the mobile yellow page service.

The LBS is facing technical and social challenges, such as location tracking, privacy issues, positioning in different environments using various locating methods, and the investment of location-aware applications.

An interesting development is the *Compose* project, which aims to overcome the drawbacks of the current solutions by pursuing a service-integrated approach that encompasses *pre-trip* and *on-trip services,* considering that on-trip services could be split into *in-car* and *last-mile services* (Bocci, 2005). The pre-trip service means the 3D navigation of the users in a city environment, and the on-trip service means the in-car and the last-mile services together. The in-car service is the location-based service and the satellite broadcasting/multicasting. In this case, the user has wireless-link access by PDA to broadcast or multicast. The last-mile service helps the mobile user with a PDA to receive guidance during the final part of the journey.

In order to create applications that utilize multicast over LBS solutions, the middleware platform of LocatioNet Systems provides all required service elements such as end-user devices, service applications, and position determination technologies, which are perfectly integrated (LocatioNet, 2006). The middleware platform of LocatioNet has an open API and a *software development kit* (SDK). Based on these, the application developers are able to easily implement novel services, focusing on comfortable user interface and free from complex details of the LBS (LocatioNet, 2006).

The article focuses on the multicast solutions over the current LBS solutions. This kind of communication is in fact a special case of the multicast communication model, called geocast, where the sender disseminates the data to a subset of the multicast group members that are in a specific geographical area.

MULTICASTING

The models of the multicast communication differ in the realization of the multiplication function in the intermediate nodes. In the case of the datalink level, the intermediate nodes are switches; in the network level, they are routers; and in the application level, the fork points are applications on hosts.

The datalink-level-based multicast is not flexible enough for new applications therefore it has no practical importance. The *network-level multicast* (NLM)—named IP-multicast—is well elaborated, and sophisticated routing protocols are developed for it. However, it has not been deployed widely yet since routing on the whole Internet has not been solved perfectly. The application-level solution gives less efficiency compared to the IP-multicast, however its deployment depends on the application itself and it has no influence on the operation of the routers. That is why the *application-level multicast* (ALM) currently has an increasing importance.

There are a lot of various protocols and implementations of the ALM for wired networks. However, the communication over wireless networks further enhances the importance of the ALM. The reason is that in the case of mobile devices, the importance of ad-hoc networks is increasing. Ad-hoc is a network that does not need any infrastructure. Such networks are *Bluetooth* (Haartsen, 1998) and *mobile ad hoc network* (MANET), which comprise a set of wireless devices that can move around freely and communicate in relaying packets on behalf of one another (Mohapatra et al., 2004).

In computer networking, there is a weaker definition of this ad-hoc network. It says that ad-hoc is a computer network that does not need routing infrastructure. It means that the mobile devices that use base stations can create an *ad-hoc computer network.* In such situations, the usage of *application-level networking* (ALN) technology is more practical than IP-multicast. In order to support this group communication, various multicast routing protocols are developed for the mobile environment. The multicast routing protocols for ad-hoc networks differ in terms of state maintenance, route topology, and other attributes.

The speed and reliability of sharing the information among the communication software entities on individual hosts depends on the network model and the topology. Theory of *peer-to-peer* (P2P) networks has gone through a great development in past years. Such networks consist of peer nodes. Usually, registered and reliable nodes connect to a grid, while P2P networks can tolerate the unreliability of nodes and the quick change of their numbers (Uppuluri, Jabisetti, Joshi, & Lee, 2005). Generally the ALN solutions use the P2P communication model, and multicast services overlay the P2P target created by the communicating entities.

The simplest ad-hoc multicast routing methods are *flooding* and *tree-based routing.* Flooding is very simple, which offers the lowest control overhead at the expense of generating high data traffic. This situation is similar to the traditional IP-multicast routing. However, in a wireless ad-hoc environment, the tree-based routing fundamentally differs from the situation in wired IP-multicast, where tree-based multicast routing algorithms are obviously the most efficient ones, such as in the *multicast open shortest path first* (MOSPF) routing protocol (Moy, 1994). Though tree-based routing generates optimally small data traffic on the overlay in the wireless ad-hoc network, the tree maintenance and updates need a lot of control traffic. That is why the two simplest methods are not scalable for large groups.

A more sophisticated ad-hoc multicast routing protocol is the *core-assisted mesh protocol* (CAMP), which belongs to the mesh-based multicast routing protocols (Garcia-Luna-Aceves & Madruga, 1999). It uses a shared mesh to support multicast routing in a dynamic ad-hoc environment. This method uses cores to limit the control traffic needed to create multicast meshes. Unlike the core-based multicast routing protocol as the traditional *protocol independent multicast-sparse mode* (PIM-SM) multicast routing protocol (Deering et al., 1996), CAMP does not require that all traffic flow through the core nodes. CAMP uses a receiver-initiated method for routers to join a multicast group. If a node wishes to join to the group, it uses a standard procedure to announce its membership. When none of its neighbors are mesh members, the node either sends a join request toward a core or attempts to reach a group member using an expanding-ring search process. Any mesh member can respond to the join request with a join *acknowledgement* (ACK) that propagates back to the request originator.

In contrast to the mesh-based routing protocols, which exploit variable topology, the so-called gossip-based multicast routing protocols exploit randomness in communication and mobility. Such multicast routing protocols apply gossip as a form of randomly controlled flooding to solve the problems of network news dissemination. This method involves member nodes talking periodically to a random subset of other members. After each round of talk, the gossipers can recover their missed multicast packets from each other (Mohapatra et al., 2004). In contrast to the deterministic approaches, this probabilistic method will better survive a highly dynamic ad hoc network because it operates independently of network topology, and its random nature fits the typical characteristics of the network.

THE LOCATION-AWARE MULTICAST

An interesting type of ad-hoc multicasting is the *geocasting.* The host that wishes to deliver packets at every node in a certain geographical area can use such a method. In such case, the position of each node with regard to the specified geocast region implicitly defines group membership. Every node is required to know its own geographical location. For this purpose they all can use the *global positioning system* (GPS). The geocasting routing method does not require any explicit join and leave actions. The members of the group tend to be clustered both geographically and topologically. The geocasting routing exploits the knowledge of location.

The geocasting can be combined with flooding. Such methods are called *forwarding zone* methods, which constrain the flooding region. The forwarding zone is a geographic area that extends from the source node to cover the geocast zone. The source node defines a forwarding zone in the header of the geocast data packet. Upon receiving a geocast packet, other machines will forward it only if their location is inside the forwarding zone. The *location-based multicast* (LBM) is an example for such *geocasting-limited flooding* (Ko & Vaidya, 2002).

LBM GEOCASTING AND IP MULTICAST

Using LBM in a network where routers are in fixed locations and their directly connected hosts are in a short distance, the location of the hosts can be approximated with the location of their router. These requirements are met by most of the GSM/UMTS, WIFI/WIMAX, and Ethernet networks, therefore a novel IP layer routing mechanism can be introduced.

This new method is a simple *geocasting-limited flooding*, extending the normal multicast RIB with the geological location of the neighbor routers. Every router should know its own location, and a routing protocol should be used to spread location information between routers. The new IP protocol is similar to the UDP protocol, but it extends it with a source location and a radius parameter. The source location parameter is automatically assigned by the first router. When a router receives a packet with empty source location, it assigns its own location to it. The radius parameter is assigned by the application itself, or it can be an administratively defined parameter in routers.

Routers forward received packets to all their neighbors except the neighbor the packet arrived from, and the neighbors outside the circle defined by the source location and radius parameters. If a packet arrives from more than one neighbor, only the first packet is handled; the duplicates are dropped.

This method requires changes in routing operating systems, but offers an easy way to start geocasting services on existing IP infrastructure without using additional positioning devices (e.g., GPS receiver) on every sender and receiver. The real advantages of the method are that geocasting services can be offered for all existing mobile phones without any additional device or infrastructure.

PRODUCT MANAGEMENT

The product managers have to take responsibility for the software and hardware part of the LBS product. This is a very important part of the process because, if the development engineer leaves a useful product out of consideration, the whole project could possibly be led astray. There are several product-management systems that help to coordinate the whole process.

FUTURE TRENDS

The multicast communication over mobile ad-hoc networks has increasing importance. This article has described the fundamental concepts and solutions. It especially focused on the area of *location-based services* (LBS) and the possible multicasting over the LBS systems. It was shown that a special kind of multicast, called *geocast* communication model, utilizes the advantages of the LBS since it is based on the location-aware information being available in the location-based solutions.

There are two known issues of this IP-level geocasting. The first problem is the scalability, the flooding type of message transfer compared to multicast tree-based protocols is less robust, but this method is more efficient in a smaller environment than using tree allocation overhead of multicast protocols. The second issue is a source must be connected directly to the router, being physically in the center position, to become source of a session. The proposed geocasting-limited flooding protocol should be extended to handle those situations where the source of a session and the target geological location are in different places.

CONCLUSION

The basically two different technologies, the location-based services in the mobile communication and the well-elaborated multicast technology, are jointed in the multicast over LBS solutions. As it was described, this emerging new area has a lot of possibilities that have not been completely utilized.

As a conclusion it can be stated that despite the earlier predicted slower development rate of the LBS solutions, nowadays the technical possibilities and the consumers' demands have already met. Furthermore, based on the latest development of the multicast over P2P technology, the one-to-many communication can be extended to the LBS systems. Also, in the case of the emerging homeland security applications, the multicast over LBS is not only a possibility, but it became a serious requirement as well. The geospatial property of the LBS provides technical conditions to apply a specialized type of the multicast technology, called geocasting, which gives an efficient and users' group targeted solution for the one-to-many communication.

REFERENCES

Bocci, L. (2005). *Compose project web site.* Retrieved from http://www.newapplication.it/compose

Chen, Y., Chen, Y. Y., Rao, F. Y., Yu, X. L., Li, Y., & Liu, D. (2004). LORE: An infrastructure to support location-aware services. *IBM Journal of Research & Development, 48*(5/6), 601-615.

Deering, S.E., Estrin, D., Farinacci, D., Jacobson, V., Liu, C-G., & Wei, L. (1996). The PIM architecture for wide-area multicast routing. *IEEE/ACM Transactions on Networking, 4*(2), 153-162.

Garcia-Luna-Aceves, J. J., & Madruga, E. L. (1999, August). The core-assisted mesh protocol. *IEEE Journal of Selected Areas in Communications,* 1380-1394.

Haartsen, J. (1998). The universal radio interface for ad hoc, wireless connectivity. *Ericsson Review, 3.* Retrieved 2004 from http://www.ericsson.com/review

Hightower, J., & Borriello, G. (2001). Location systems for ubiquitous computing. *IEEE Computer,* (August), 57-65.

Hofmann-Wellenhof, B., Lichtenegger, H., & Collins, J. (1997). *Global positioning system: Theory and practice* (4th ed.). Vienna/New York: Springer-Verlag.

Hosszú, G. (2005). Current multicast technology. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and information technology* (pp. 660-667). Hershey, PA: Idea Group Reference.

Ibach, P., Tamm, G., & Horbank, M. (2005). Dynamic value webs in mobile environments using adaptive location-based services. *Proceedings of the 38th Hawaii International Conference on System Sciences.*

Ibach, P., & Horbank, M. (2004, May 13-14). Highly-available location-based services in mobile environments. *Proceedings of the International Service Availability Symposium,* Munich, Germany.

Ko, Y-B., & Vaidya, N. H. (2002). Flooding-based geocasting protocols for mobile ad hoc networks. *Proceedings of the Mobile Networks and Applications, 7*(6), 471-480.

LocatioNet. (2006). *LocatioNet and Ericsson enter into global distribution agreement.* Retrieved from http://www.locationet.com

Mohapatra, P., Gui, C., & Li, J. (2004). Group communications in mobile ad hoc networks. *Computer, 37*(2), 52-59.

Moy, J. (1994, March). *Multicast extensions to OSPF.* Network Working Group RFC 1584.

Niedzwiadek, H. (2002). *Location-based services for homeland security.* Retrieved March 2006 from http://www.jlocationservices.com/LBSArticles

Uppuluri, P., Jabisetti, N., Joshi, U., & Lee, Y. (2005, July 11-15). P2P grid: Service oriented framework for distributed resource management. *Proceedings of the IEEE International Conference on Web Services,* Orlando, FL.

KEY TERMS

Ad-Hoc Computer Network: Mobile devices that require base stations can create the ad-hoc computer network if they do not need routing infrastructure.

Ad-Hoc Network: A network that does not need any infrastructure. One example is Bluetooth.

Application-Level Multicast (ALM): A novel multicast technology that does not require any additional protocol in the network routers, since it uses the traditional unicast IP transmission.

Application-Level Network (ALN): The applications, which are running in the hosts, can create a virtual network from their logical connections. This is also called overlay network. The operations of such software entities are not able to understand without knowing their logical relations. In most cases these ALN software entities use the *P2P model,* not the *client/server* for the communication.

Client/Server Model: A communicating method, where one hardware or software entity (server) has more functionalities than the other entity (the client), whereas the client is responsible to initiate and close the communication session towards the server. Usually the server provides services that the client can request from the server. Its alternative is the *P2P model.*

Geocast: One-to-many communication method among communicating entities, where an entity in the root of the multicast distribution tree sends data to that certain subset of the entities in the multicast dissemination tree, which are in a specific geographical area.

Multicast: One-to-many and many-to-many communication method among communicating entities on various networked hosts.

Multicast Routing Protocol: In order to forward the multicast packets, the routers have to create multicast routing tables using multicast routing protocols.

Peer-to-Peer (P2P): A communication method where each node has the same authority and communication capability. The nodes create a virtual network, overlaid on the Internet. The members organize themselves into a topology for data transmission.

Product Management: A function within a corporation dealing with the continuous management and welfare of the products at all stages of the production procedure in order to ensure that the products profitably meet the needs of customers.

M-Advertising

Michael Decker
University of Karlsruhe, Germany

INTRODUCTION

According to our comprehension, mobile advertising (also called “wireless advertising” or “mobile marketing”) is the presentation of advertising information on mobile handheld devices with a wireless data link like cellular phones, personal digital assistants and smartphones; however notebooks/laptops and tablet PCs are not considered as mobile devices in this sense, because they are used like stationary devices at different locations. For example SMS-messages with product offers would be a simple form of m-advertising. In this article we discuss the special features of m-advertising, but also the problems involved. Afterwards we name basic methods of m-advertising and compare their general strengths and weaknesses using a set of criteria.

M-ADVERTISING COMPARED TO TRADITIONAL FORMS OF ADVERTISING AND INTERNET-BASED ADVERTISING

Conventional media for advertising are newspapers, advertising pillars, TV and radio commercials. Relative new media for advertising are the Internet and mobile devices. Both have some features in common:

- **Individually Addressable:** The user can be addressed individually, so a high degree of personalization is possible: it is possible to tailor the content of each advert according to the profile of the consumer (mass customization).
- **Interactive:** if end users receive an advert, they can immediately request further information, participate in a sweepstake or forward an advert to friends. The last one is especially interesting in terms of “viral marketing.”
- **Multimedia-Capable:** Multimedia elements (e.g., pictures, movies, jingles, tunes, sounds) are important to realize entertaining adverts and to generate brand awareness.
- **Countable:** Each impression of an ad can be counted; for most conventional methods of advertising like TV/radio/cinema commercials or adverts in print media this can not be done and thus the advertisers are billed according to a rough estimate of the number of generated contacts.

But there are additional features of m-advertising [see also Barnes (2002)]:

- **Context:** In the sense of mobile computing, context is “[…] any information that can be used to characterize the situation of an entity” (Dey, 2001). This information helps to support a user during an interaction with an application. For mobile terminals with their limited user interface context-awareness is especially important. The most prominent example of context is the location of a user, because it changes often and there are a lot of useful scenarios of how to exploit that information. The location information for these “location-based services” can be retrieved based on the position of the currently used base station, the runtime difference when using more than two base stations (TDOA: Time Differential of Arrival) or using a GPS-receiver (Zeimpekis, Giaglis, & Lekakos, 2003). Other examples of context information also used for m-advertising are “weather” and “time” (Salo & Tähtinen, 2005).
- **Reachability:** People carry their mobile terminal along with them most of the day and rarely lend it away or share it with other people, because it is a personal device. Therefore marketers can reach people almost anywhere and anytime.
- **Convenience:** Mobile terminals are much simpler to handle than personal computers because they are preconfigured by the mobile network operator and have no boot-up time, so they are a medium for electronic advertising to reach people who don’t want to use a computer.
- **Penetration Rates:** Mobile terminals—especially cellular phones—have very high penetration rates, which exceed those of fixed line telephones and personal computers. Mobile terminals are more popular than PCs because they are more affordable and simpler to handle. The current global number of cellular phones is beyond one billion, there are even countries with penetration rates over 100 % (Netsize, 2005).

At first glance, m-advertising seems to be a direct continuation of Internet-based advertising: instead of a fixed

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

computer with a wired data link a mobile terminal with wireless data link is used. But most forms of advertising in the Internet are based on the idea of showing additional advertising information on the user interface (banners on Web pages, sponsored links in search engines). Due to the limited display size of mobile terminals these forms of advertising can't be used for m-advertising, so new methods have to be developed.

Mobile terminals are much more personal devices than personal computers, so a higher degree of personalization can be obtained than with Internet advertising, which leads to better response rates than with other forms of direct marketing (Kavassalis et al., 2003).

CHALLENGES

As shown in the last section, m-advertising has some unique advantages when compared to other forms of advertising. But one shouldn't conceal the challenges associated with it:

- **Unsolicited Messages:** Unsolicited mass-mailing with commercial intention ("Spam") as well as malware (viruses, trojan horses, spyware, etc.) are a great worriment in the fixed-line Internet; the portion of spam messages in e-mail communication exceeds the 50% mark by far. Unsolicited messages on mobile terminals are a much bigger problem, because mobile terminals have limited resources to handle them and are personal devices.
- **Limited Usability:** Due to the limited dimensions of mobile terminals they have only a small display and no real keyboard. This has to be considered when designing adverts for mobile terminals. One cannot ask the user for extensive data input, for example, about his/her fields of interests or socio-demographic particulars.
- **Limited Resources:** Mobile terminals have very limited resources, for example, memory, CPU-power and available bandwidth. These have to be considered when designing adverts for mobile terminals, for example, transmission of adverts with a lot of data volume is not adequate,
- **Privacy Concerns:** As already mentioned, mobile terminals are personal devices with personal data stored on them; it is also possible to track the location of the users.
- **Cost of Mobile Data Communication:** Mobile data communication is still very expensive in some regions, so no consumer wants to cover the costs caused by the transmission of adverts.
- **Technical Heterogeneity:** The underlying network infrastructure and the capabilities of mobile terminals are much more heterogeneous than for ordinary fixed-line computers: one advert might look great on one type of terminal, but isn't displayable on another one. It might cause significant costs when the creator of an m-advertising campaign has to consider many different types of mobile terminals.

Due to the problems with unsolicited e-mails and telephone calls "permission-based marketing" is a generally accepted principle when designing systems and campaigns for m-advertising (Barwise & Strong, 2002): A consumer will only receive advertising-messages on his mobile device if he explicitly gave his permission. The adverts sent to a user will be chosen according to the profile of interests of the user. But there is one problem with this principle: a consumer has to know about an m-advertising campaign or system to give his explicit agreement, so one has to "advertise for m-advertising." Because of this m-advertising is very often integrated into bigger campaigns along with traditional media, see Kavassalis et al. (2003) for examples.

DIFFERENT APPROACHES FOR M-ADVERTISING

We distinguish different approaches for m-advertising by the underlying mode of wireless communication used:

- When using broadcast communication, messages are sent to all ready-to-receive terminals in the area covered by the radio waves according to their natural propagation. If the area covered is rather restricted we talk about a local broadcast, which allows realizing a certain degree of location-aware adoption; the opposite case is global broadcast. Examples: cell-broadcast to all terminals in a certain network-cell (local broadcast) or digital television standards like DVB-H or DMB (global broadcast).
- Mobile ad-hoc networks (MANETs) are wireless networks without a dedicated infrastructure or a central administration (Murthy & Manoj, 2004). Two terminals of such a network exchange messages when the distance between them is short enough, a message can also be routed via several terminals to the recipient (multi-hop).
- Unicast communication provides a dedicated point-to-point connection between a base station and a mobile terminal in an infrastructure-based network like GSM, WLAN or UMTS.

To compare different advertising approaches based on these modes we apply the following set of criteria:

- Which degree of personalization and location-aware adoption can be achieved?

- Will there be costs for the data transmission which have to be borne by the end-user?
- How reliable is the transmission of the advert? Can it be guaranteed that the advert is received indeed within a certain time span? The last point might be crucial for campaigns with temporary and "last minute" offers.
- Is the mode capable of direct user interaction?
- Unsolicited messages aren't bearable on mobile devices, so a system for m-advertising should be designed to make the dispatching of unsolicited messages impossible.
- Is it possible to detect the exact number of contacts generated?

BROADCASTING

Using broadcast communication, adverts are addressed undirected to an anonymous crowd of people, so there is no possibility of personalization. If the broadcast is a local one at least a certain degree of location-aware adoption can be achieved, for example adverts about shopping facilities and sights located in one cell of a GSM-network. Receiving broadcast messages is free of charge for the end-users, but if the user hasn't turned on his device or isn't in the area covered by the broadcast he may miss a message. If a user wants to respond to an advert or forward it to other people he has to resort to unicast communication. Unsolicited messages can only occur if the user is in ready-to-receive state.

Using broadcasting-based advertising on mobile terminals we lose a lot of the specific opportunities of m-advertising, for example, the high personalization and the interactive nature, but there are no costs for data transmission for the end-user and his anonymity is protected. Without further measures it is not possible to find out how many users received an advertisement.

MANETs

The idea of MANETs can be applied for the distribution of adverts (Ratsimor, Finin, Joshi, & Yesha, 2003; Straub & Heinemann, 2004): following the idea of "word-of-mouth" recommendations mobile terminals exchange adverts when they approach each other, whereas the initial transmission of adverts is performed by fixed "information sprinklers." If these are positioned at an appropriate place we can obtain a certain degree of location awareness at least for those users whose devices receive the message from the sprinklers directly. A client application installed on the mobile terminal will only display adverts that match the profile of the user, so a certain degree of personalization is possible. There are also incentive schemes: when an advert leads to a purchase a user whose device participated in the recommendation chain may receive a bonus.

The multi-hop case requires a special client application on the device. At present only the single hop-case can be found in practice: "sprinklers" installed at appropriate places (e.g., entrance halls of shopping or conference centers or even at billboards) submit adverts to mobile devices using infrared [e.g., "marketEye™" by Accinity (2006)] or bluetooth [e.g., "BlipZones ™" by BLIP Systems (2006)] communication, but the adverts are not transmitted to other devices automatically.

M-advertising based on MANETs doesn't cause costs for the end-user. The reliability of this form of advertising is not very high, because one cannot give guarantees how long it takes until a consumer receives—if at all—an ad. For interaction the user has to resort to Unicast communication. In the multi-hop case the local client applications can decide which adverts to display and which not to, so the user won't be harassed by unsolicited messages; in the single-hop case the user can disable his infrared or bluetooth interface if he doesn't want to receive ads. The advertisers can only count contacts that led to certain defined actions (e.g., if a digital voucher is redeemed). Receiving ads from unknown mobile terminals all the time may also be very energy consuming and there is the danger of receiving malware which exploits flaws of the mobile device.

UNICAST

Unicast-communication can be further divided into "push" and "pull:" in push-mode the consumer receives a message without a direct request, for example, SMS; in pull-mode the consumer has to perform an explicit request for each message, for example, request of WAP-pages. Since Unicast communication provides a dedicated point-to-point connection, the advertiser can deliver a different ad for each user, so there is a high degree of personalization possible and the advertiser can calculate the exact number of contacts generated.

The most obvious form of pull-advertising for mobile terminals is Web-pages in special formats like WML or cHTML. When viewing such pages the user has to pay for the data volume, so there is the idea that the advertiser covers the costs if the user's profile meets certain criteria (Figge, Schrott, Muntermann, & Rannenberg, 2002). Pull-advertising isn't vulnerable for unsolicited messages, but the reliability is also restricted, because the user might miss offers he is interested in or obtain offers too late if he doesn't request the right page at the right time.

SMS as form of mobile push-messaging is also the most successful data service for mobile phones, but the messages can only consist of text and are bound to a maximum size of 140 bytes or 160 letters for a 7-bit-encoding—and Barwise and Strong (2002) even recommend to use much shorter messages when using SMS for advertising. But it is the data service which is supported by almost all cellular phones, so it is an interesting opportunity for m-advertising.

The multimedia messaging service (MMS) is a further development of SMS, which is also capable of displaying multimedia content, but the creation of an appealing message on a mobile device is challenging; also in many countries sending MMS is very expensive. "V-Card" (Mohr, Nösekabel, & Keber, 2003) is a platform for the creation of MMS and offers templates and multimedia content, so the user can design easily a personalized message. To keep the service free of charge for the end-user a sponsor bears the costs and thus has his advertising message included in the MMS.

The "MoMa" system (Bulander, Decker, Schiefer, & Kölmel, 2005) is a system for context aware push advertising for mobile terminals with a special focus on privacy aspects. The system acts as mediator between end-users and advertisers: end-users put "orders" into the system using a special client application to express that they are interested in advertising concerning a certain product or service (e.g., restaurants or shoe shops in their current surrounding); the possible products and services are listed in a hierarchical catalogue. Depending on their type the orders can be refined through the specification of attributes. On the other side of the system the advertisers put "offers" into the system. A matching component tries to find fitting pairs of orders and offers, in case of a hit a notification via SMS/MMS, e-mail or text-to-speech call will be dispatched. When appropriate the matching process also considers context-parameters like the location of the user or the weather.

Push advertising in general provides a high degree of reliability but is also vulnerable with regard to unsolicited messages.

CONCLUSION

We highlighted the special features of mobile and wireless terminals as medium for advertising, but saw also the considerable challenges. M-advertising might be one of the first successful m-business applications in the business-to-consumer sector, because it doesn't require m-payment. The current market share of m-advertising in relation to the total advertising market seems to be less than one percent in most countries, but is expected to reach a few percent-points (like nowadays Internet-advertising) in the medium term.

REFERENCES

Accinity. (n.d.). Retrieved March 28, 2006, from http://www.accinity.com

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing*, *16*(1), 14-24.

Barnes, J. (2002). Wireless digital advertising: Nature and implications. *International Journal of Advertising*, *21*(3), 399-420.

BLIP Systems. (n.d.). Retrieved March 28, 2006, from http://www.blipsystems.com

Bulander, R., Decker, M., Schiefer, G., & Kölmel, B. (2005). Advertising via mobile terminals. In *Proceedings of the 2nd International Conference on E-Business and Telecommunication Networks (ICETE '05)*. Reading, UK.

Dey, A.K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal*, *5*(1), 4-7.

Figge, S., Schrott, G., Muntermann, J., & Rannenberg, K. (2002). Earning M-Oney—A situation based approach for mobile business models. In *Proceedings of the 11th European Conference on Information Systems (ECIS)*. Naples, Italy.

Kavassalis, P., Spyropoulou, N., Drossos, D., Mitrokostas, E., Gikas, G., & Hatzistamatiou, A. (2003). Mobile permission marketing: Framing the market inquiry. *International Journal of Electronic Commerce*, *8*(1), 55-79.

Mohr, R., Nösekabel, H., & Keber, T. (2003). V-Card: Sublimated message and lifestyle services for the mobile mass market. In *Proceedings of the 5th International Conference on Information and Web-Based Applications and Services*. Jakarta, Indonesia.

Murthy, C., & Manoj, B. (2004). *Ad hoc wireless networks—Architectures and protocols*. Upper Saddle River, NJ: Prentice Hall.

Netsize. (2005). *The Netsize guide 2005*. Paris, France.

Ratsimor, O., Finin, T., Joshi, A., & Yesha, Y. (2003). eNcentive: A framework for intelligent marketing in mobile peer-to-peer environments. In *Proceedings of the 5th ACM International Conference on Electronic Commerce* (ICEC 2003). Pittsburgh, Pennsylvania.

Salo, J., & Tähtinen, J. (2005). Retailer use of permission-based mobile advertising. In I. Clarke, III & T.B. Flaherty (Eds.), *Advances in Electronic Marketing* (pp. 139-155). Hershey, PA: Idea Group Publishing.

Straub, T., & Heinemann, A. (2004). An anonymous bonus point system for mobile commerce based on word-of-mouth-recommendation. In *Proceedings of the ACM Symposium on Applied Computing (SAC '04)*. Nicosia, Cyprus.

Zeimpekis, V., Giaglis, G., & Lekakos, G. (2003). A taxonomy of indoor and outdoor positioning techniques for mobile location services. *ACM SIGecom Exchanges*, *3*(4), 19-27.

KEY TERMS

Broadcast: sending data using wireless communication to an anonymous audience according to the natural propagation of radio waves.

Context: information available at runtime of a computer system to support the user when interacting with the system.

Location-Based Services: most prominent case of context-aware services, which adapt themselves according to the current location of the user.

Mobile Advertising (M-Advertising): adverts displayed on mobile and wireless terminals like cellular phones, smartphones or PDAs.

Mobile Ad-Hoc Network (MANET): Wireless network without infrastructure and central administration, the nodes (devices) pass messages to other nodes in reach.

Pull: A user receives a message only as direct response.

Push: A user receives a message without directly requesting it.

Unicast: Communication using an infrastructure-based network which provides dedicated point-to-point-connection. Examples are GSM/GPRS, WLAN or UMTS.

Viral Marketing: Dissemination of adverts by consumers themselves ("tell a friend!"); is expected to generate exponential rates of contacts and a higher trustability than adverts received from firms directly.

Man–Machine Interface with Applications in Mobile Robotic Systems

Milan Kvasnica
Tomas Bata University, Zlin, Czech Republic

INTRODUCTION

This article focuses on the current state-of-the-art assistive technologies in man-machine interface and its applications in robotics. This work presents the assistive technologies developed specifically for disabled people. The presented devices are as follows:

- The head joystick works on a set of instructions derived from intended head movements. Five laser diodes are attached to the head at specific points whose light rays' spots are scanned by a set of CCD cameras mounted at strategic locations (on the ceiling, on the wall, or on a wheelchair).
- Automatic parking equipment has two laser diodes attached at the back of the wheelchair, and their light rays' spots are scanned by the CCD cameras.
- A range-inclination tracer for positioning and control of a wheelchair works on two laser diodes attached onto the front of the wheelchair. A CCD camera mounted on the front of the wheelchair detects the light rays' spots on the wall.
- The body motion control system is based on a set of instructions derived from intended body motion detected by a six component force-torque transducer, which is inserted between the saddle and the chassis of the wheelchair.
- An optoelectronic handy navigator for blind people consists of four laser diodes, the 1-D CCD array (alternatively PSD array), a microprocessor, and a tuned pitch and timbre sound source. The functionality of this system is based on the shape analysis of the structured lighting. The structured lighting provides a cutting plane intersection of an object, and a simple expert system can be devised to help blind people in classification and articulation of 3-D objects. There are two parameters involved: the distance and the inclination of the object's articulation. The time-profiles of the distance and inclination are used to adjust the frequency and amplitude of the sound generator. The sound representation of a 3-D object's articulation enables the skill-based training of a user in recognizing the distance and ambient articulation.

The head joystick, the automatic parking equipment, the range-inclination tracer, and the body motion control system for the wheelchair control are suitable for people who have lost the ability to use their own lower limbs to walk or their upper limbs for quadriplegics. The optoelectronic handy navigator is suitable for blind people. The mentioned sensory systems help them to perform daily living tasks, namely to manage independent mobility of electrical wheelchair or to control a robot manipulator to handle utensils and other objects. The customization of described universal portable modules and their combinations enable convenient implementation in rooms and along corridors, for the comfort of the wheelchair user. Smart configuration of the optoelectronic handy navigator for blind people enables the built-in customization into a handy phone, handheld device, or a white stick for blind people.

BACKGROUND

Significant progress in human-computer interfaces for elderly and disabled people has been reported in recent years. Some examples for such devices are the eye-mouse tracking system, hand gesture systems, face gesture systems, head controller, head joystick, and human-robot shoulder interface, all presented at recent international conferences. The aim of this article is to publish further progress in the field of assistive technologies like the head joystick, automatic parking equipment, range-inclination tracer, body motion control system, and optoelectronic handy navigator for blind people.

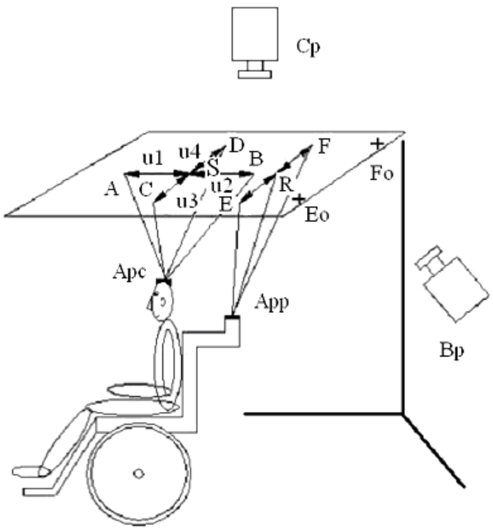
HEAD JOYSTICK AND AUTOMATIC PARKING EQUIPMENT

Following parts of the modular sensory system enables the processing of multi-DOF information for the control and the positioning of a wheelchair by means of two types of modules for alternative use, as shown in Figure 1.

The module Apc (the module of four laser diodes) is designed for tracking the head motion of the wheelchair user. The ceiling-mounted CCD cameras detect the Apc laser rays. The fifth, auxiliary laser diode with redundant light

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. The modules Apc and App positioned relative to the plain of the ceiling and the module Bp of the CCD camera mounted in perpendicular view or in perspective view against the light spots on the ceiling



spot is used for the verification of accurate functionality. The module App (the module of two laser diodes) is designed for automatic parking of the wheelchair into a predefined position in the room. The third auxiliary laser diode with redundant light spot is used for the verification of accurate functionality. The modules Apc and App have the presetting control 1 of the angle 2s contained by mutual opposite light rays 2, as shown in Figures 1 and 2. The auxiliary fifth or third laser diode is centered in the axis. The camera with 2-D CCD array can be arranged in two ways:

- The perpendicular view downwards against the translucent screen, which is mounted parallel to the ceiling. The module Cp for direct sampling is shown in Figure 3. The light spots reflected by light rays of Apc, or App respectively, are sampled by the camera with a 2-D CCD array. The module Cp for direct sampling of the light spots (no. 3) consists of the camera with 2-D CCD array (no. 6) with focusing optics (no. 5) and the flange (no. 1) mounted perpendicular to the translucent screen. The light spots from the laser light rays are projected onto the translucent screen (no. 4). The translucent screen spans the entire ceiling of the room. In larger rooms, four Cp modules are attached onto the ceiling in front of the laser rays.
- The perspective view of the ceiling and light spots of the laser ray images spots from the modules Apc and App are shown in Figure 1. The module Bp, depicted in Figure 4, is shown in Figure 1 in perspective view on the wall. The camera with 2-D array makes the sampling of the light spot position from the laser light rays on the ceiling plane. The X-Y coordinate system on the ceiling is used to monitor the parking position of the wheelchair.

The module Cp is mounted on the ceiling against the modules Apc and App, respectively. The Apc module is attached to the head of the wheelchair user, and the rays are intersecting the ceiling plain of the Bp or Cp modules respectively, in light spots A, B, C, and D. The intersection of abscises AB and CD is the point S centered by auxiliary laser diode. The lengths of abscises AS, SB, CS, and SD are u_1, u_2, u_3, and u_4. The App module is attached to the wheelchair, and the light rays intersect the ceiling plain in light spots E, F in equal distance u from the middle point R centered by auxiliary laser diode. The light spots position and configuration is analyzed and processed for the navigation of a wheelchair.

Purposeful head motion of the module Apc represented by the light spots configuration is sampled by means of the modules Bp, or Cp from the ceiling. The following commands are used for three degrees-of-freedom control of the wheelchair with an adjustable operating height of the wheelchair perpendicular to the 2-D coordinate frame on the ground:

- **Start:** The head movement with the module Apc outwards the dead zone position of the light spots.
- **Stop:** The head movement with the module Apc inwards the dead zone position of the light spots.
- The dead zone is defined by the ratio u_1/u_2 and $u_3/u_{4,}$ for example (only for perpendicular view of the CCD Camera) by the interval <0,8; 1,2> and for the angle Ω by the interval $<-15^0;+15^0>$. Inside these intervals, following commands are not valid because of physiological trembling of the head. Outwards these intervals are valid following commands.
- **Forward, Backward for the First DOF:** The dividing ratio of diagonals $u_1/u_2 > 1{,}2$ respectively $u_1/u_2 < 0{,}8$.
- **Up, Down for the Second DOF:** The dividing ratio of diagonals $u_3/u_4 > 1{,}2$ respectively $u_3/u_4 < 0{,}8$.
- **Turn to the Left – Turn to the Right for the Third DOF:** Last increment of the angle Ω is positive $\Omega > +15^0$, respectively, and negative $\Omega < -15^0$ oriented.
- The magnitude of the dividing ratios u_1/u_2, u_3/u_4, and the angle Ω is assigned to the velocity of the wheelchair motion for each DOF.
- The motion control system of the wheelchair enables parallel control of all three degrees-of-freedom.
- The light spot A from the module Apc is recognized by means of the enhanced intensity, color, or shape in contrast with light spots B, C, and D. This is needed for the orientation of the wheelchair against the basic light spot position.

Figure 2. The multi-laser configuration modules App, and Apc

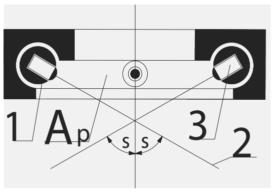


Figure 3. The Cp module for direct sampling

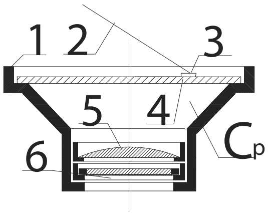


Figure 4. The module Bp for the ground plane or for perspective ceiling sampling

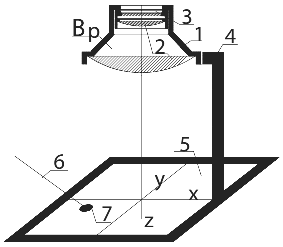
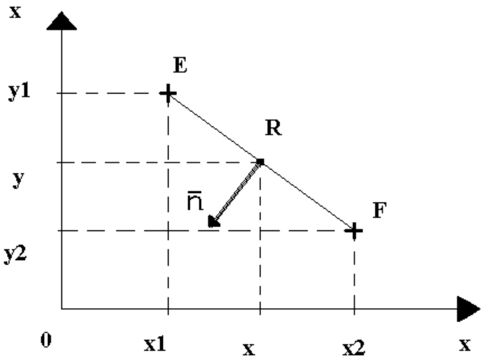


Figure 5. Navigation by automatic parking system



The automatic parking equipment of a wheelchair is devised on the module App, which is attached at the back of the wheelchair:

- The module App is used for automatic return into the parking position using controls in forward-backward, up-down, and turns right-left.
- The module App and the module Apc operate independently and sequentially into common modules Bp or Cp.
- The position E(x_1,y_1) and F(x_2,y_2) of the light spots from the module App on the ceiling are sampled by the camera. The information about the light spots position E(x_1,y_1) and F(x_2,y_2) is sufficient for the navigation into the points E_0(x_{01},y_{01}) and F_0(x_{02},y_{02}) of the basic position highlighted from the reflexive material on the ceiling.
- The light spot E from the module App is recognized by means of the intensity, color, or shape in contrast with light spot F. This is needed for the orientation of the wheelchair against the basic light spot parking position.
- Another technique of recognizing the orientation is to sample every light spot in separate picture synchronized with sequentially switched light rays.

The coordinates of the position x,y of the wheelchair at point R of the ceiling rectangular coordinate frame is given by the relationship (1), and the normal vector n in the middle point R determines the direction of the wheelchair trajectory.

$$x = \frac{1}{2}(x_1 + x_2); \quad y = \frac{1}{2}(y_1 + y_2); \tag{1}$$

- The direction **n** of the wheelchair motion in the ceiling rectangular coordinate frame is given by the relationship (3) according to Figure 5.
- All coordinates x_i, y_i, z are with respect to the geometrical center of the light spots.
- The z coordinate is used for calibration of the magnitude x_i, y_i according to the position of the light spot image center of the 2-D CCD coordinate frame.

The distance z of the wheelchair from the ceiling in the rectangular coordinate frame (recommended the value s = arctg(1/2)) according to Figures 2 and 5 is given by the relationship (2).

$$z = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}; \tag{2}$$

$$n = -\left[\frac{y_2 - y_1}{x_2 - x_1}\right]^{-1}. \tag{3}$$

The coordinates of the wheelchair position x,y and the normal vector direction n of the wheelchair motion are used for the planning of the wheelchair trajectory as well as for collision avoidance.

THE RANGE-INCLINATION TRACER

The range-inclination tracer consists of the camera with 1-D CCD array or linear PSD element and two laser diodes radiating two intersecting light rays against the wall, as shown in Figures 6 and 7. Two light spot positions of the light rays are sampled by a simple sampling algorithm, in order to evaluate the distance D and the inclination β of the wall against the wheelchair. The sampling of every light spot coordinate in separate picture synchronized with sequentially switched light rays enables recognition of the varying orientation of the light spots before and after the crossing point of the laser light rays.

Control algorithms are derived in the rectangular coordinate system x,y with an origin in zero point of mutual intersection of light rays (no. 4), so that the y coordinate fuses with optical axis (no. 5) of the camera (no. 3) with 1-D CCD array, and its positive orientation leads into the camera (see Figure 7). The coordinates of the light spot on the wall (no. 6) emitted by the light source (no. 1) are designated by $A(r_1,s_1)$, and the coordinates of the light spot emitted by the light source (no.,2) are designated by $B(r_2,s_2)$. Every light ray creates the angle α/2 with optical axis of the camera. The magnitude of the coordinates r_1,r_2 is derived according to the calibration of the light spot's position in the image coordinate frame of the 1-D CCD array. Coordinates s_1,s_2 are computed from the relationship (4).

Figure 6. Range-inclination tracer attached to the wheelchair

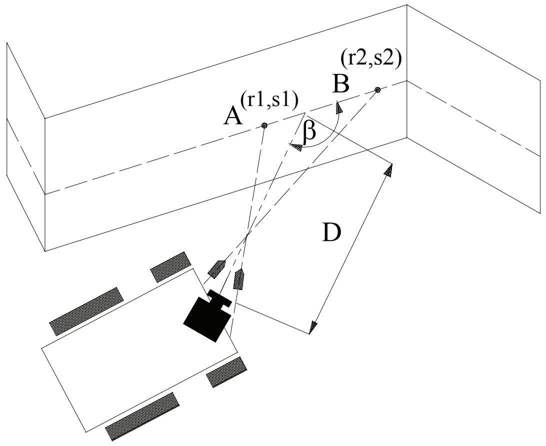
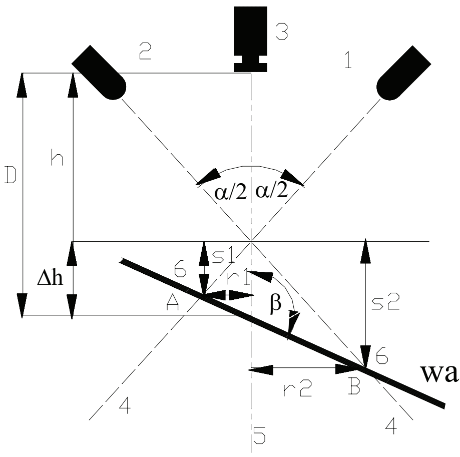


Figure 7. Geometrical approach of the activity of the range-inclination tracer



$$s_1 = \frac{r_1}{tg\frac{\alpha}{2}} \; ; \qquad s_2 = \frac{r_2}{tg\frac{\alpha}{2}} \qquad (4)$$

where h is the distance between the camera and the point of the intersection of light rays. The Δh is the difference between the point of the light rays' intersection and the intersection of the optical axis of the camera with the wall, given by the relationship (5).

The distance D between the camera and the wall is given by the relationship (6), and the inclination β of the wall against the wheelchair is given by the relationship (7).

$$D = h \pm \Delta h \qquad (5)$$

$$\Delta h = -\frac{s_1 - s_2}{r_2 - r_1} r_1 + s_1 \qquad (6)$$

$$\beta = arcctg \frac{r_2 - r_1}{s_1 - s_2} \qquad (7)$$

The described sensory system is alternatively used for independent control of the wheelchair trajectory in the vicinity of the walls, namely in long corridors, and enables the implementation of a low-cost collision avoidance system.

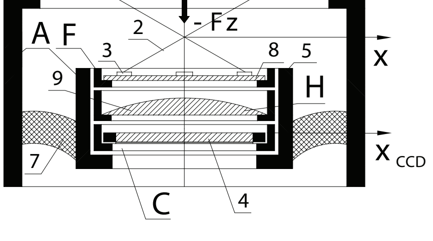
HUMAN-ROBOT INTERFACE USING THE BODY MOTION

Many people with limited mobility prefer to enjoy life. They particularly enjoy a wheelchair ride at a popular tourist spot with typical leisure activities, and sometimes go on high adventures and wilderness expeditions. These expeditions are usually long trips, and active compliant assistance by way of feedback control based on the damping device with a shock absorber is required. This active compliant assistance is able to overcome some barriers and to predict hazardous scenarios in unexpected situations like the quick halt on the stone or the raid on a sharp slope. A basic part of compliant assistance is the six-component force-torque transducer inserted between the saddle and the chassis as shown in Figures 8 and 9. The force-torque transducer used for dynamic weighing of the user and the load on the wheelchair improves the dynamic stability at maneuvering and the comfort of the ride. In addition this force-torque transducer is possible to use for the sampling of the information about the user's body motion.

The control system based on the user's body motion is based on a set of instructions derived from the body inclination and sampled by means of the six-component force-torque transducer inserted between the saddle and the chassis of the wheelchair. The six-component force-torque transducer is also used for the active compliant assistance.

The explanation of the activity is introduced on a simple modification of the six-component force-torque transducer. An example of the six-component force-torque transducer with the acting force -Fz is depicted in Figure 8. Laser diodes (no. 1) emit intersecting light rays (no. 2) creating the edges of a pyramid, intersecting the 2-D CCD array (no. 4) in light spots (no. 3). The beginning of the 3D rectangular pyramid coordinate frame x,y,z is chosen in the crossing point of the light rays (the apex of a pyramid shape). Unique light spots configuration on the 2-D CCD array changes under the force-torque acting between the flanges (no. 5 and no. 6) connected by means of elastic deformable medium (no. 7). The beginning of a floating 2D coordinates frame x_{CCD},y_{CCD} is chosen in the geometrical center of the 2-D CCD array. An even number of four light rays simplifies and enhances the accuracy of the algorithm for the evaluation of three axial shiftings and three angular displacements. In addition it removes the ambiguity of the imagination at the rotation of the 2-D CCD array around the straight line passing through two light spots, where for two inclines of opposite orientation belongs one position of the third light spot.

Figure 8. Six-component force-torque transducer



The module of five laser diodes is attached on the outer flange of the force-torque sensor, and the rays intersect the translucent screen in light spots A, B, C, and D. The intersection of the abscises AB and CD is the point S. The length of the abscises AS, SB, CS, and SD is u_1, u_2, u_3, and u_4. The light spot configuration is analyzed and processed for the navigation and positioning of the wheelchair.

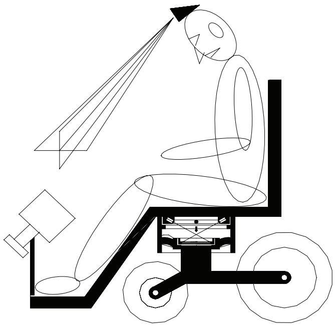
Intentional body inclinations against the saddle are represented by the light spots (no. 3) configuration in the 2-D CCD array (no. 4) of the force-torque transducer, as shown in Figure 8, which is analyzed like the force-torque acting in six components. The following commands are used for two-degrees-of-freedom control of the wheelchair as shown in Figure 9:

- **Start:** The body inclination in chosen direction outwards the light spots dead zone position.
- **Stop:** The body inclination inwards the light spots dead zone position.
- The dead zone is defined for the dividing ratio u_1/u_2 for example by the interval <0.8; 1.2>, u_3/u_4 is symmetric <0.8; 1.2>.
- **Forwards, Backwards for the First DOF:** The dividing ratio of diagonals $u_1/u_2 > 1.2$ respectively $u_1/u_2 < 0.8$.
- **Turn to the Left—Turn to the Right:** Similar for the u_3/u_4. Final direction of the wheelchair motion is the vector sum of components u_1/u_2 and u_3/u_4.
- The magnitude of the dividing ratio u_1/u_2, u_3/u_4 is assigned to the velocity of the wheelchair motion for two degrees of freedom (DOF).
- The motion control system of the wheelchair enables parallel control of two degrees-of-freedom.
- The light spot A is oriented in front direction. This is needed for the orientation of the wheelchair against the basic light spot position.

Some applications of the signal filtering for the elimination of the body inclination are needed for users with muscular trembling at neurological diseases and at the motion of the wheelchair on rough surface.

Figure 9 depicts the CCD camera for the sampling of the head joystick's light spots configuration on the floor. The difference between the sampling from the CCD camera attached to the wall or to the ceiling against the one attached to the wheelchair is in the level of navigation. The CCD camera attached to the wall or to the ceiling enables the positioning of the wheelchair with respect to the shape of the room. The CCD camera attached to the wheelchair enables only the relative positioning of the wheelchair.

Figure 9. The force-torque transducer inserted between the saddle and the chassis of the wheelchair and the CCD camera for the sampling of the head joystick's light spots configuration on the floor



OPTOELECTRONIC HANDY NAVIGATOR FOR BLIND PEOPLE

Safe and effective mobility of blind people depends largely upon reliable orientation about the articulation of the ambient. The current situation of people who suffer severe visual impairments is that they mostly require active assistance from relatives or occasionally assistance of passing by strangers in order to travel or to use transportation or to manage street traffic in crowded areas. Nevertheless, blind people, just like sighted people, do not like to be dependent on others. There are restricted sources of sound information at the traffic light control or acceptable spoken information at landmarks and significant places in information stations. Independent travel and transportation for blind people involves orienting oneself and finding a safe path through known and unknown articulated environments. Most efforts have been to solve the mobility part of the problem to help the blind traveler detect irregularities on the floor such as boundaries, objects located near or alongside his or her path in order to avoid collisions, and steer a straight and safe course through the immediate environment. Ultrasonic and laser devices for the navigation using active methods for the identification of the environment have been reported.

One way for the classification of 3-D object articulation destined for blind people is based on the principle of the twofold range-inclination tracer. The range-inclination tracer enables the shape analysis by means of the structured light-cutting plane intersection with an object. This navigation system consists of four laser diodes with focusing optics used for the light spots imagination 1-D CCD array (alternatively PSD array), microprocessor, and tuned pitch and timbre of sound source. Two parameters on the intersection are computed and evaluated: the distance and the inclination of the object.

The time-profiles of the distance and inclination, and their combinations are used to tune the sound generator's frequency and intensity. The sound representation of a 3-D object's articulation enables the skill-based training of the user in recognizing the distance and ambient articulation. Smart configuration enables the customizing of navigation device into:

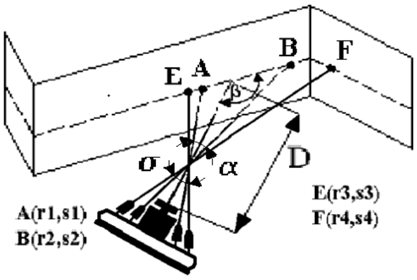
- a handy phone that uses only one line from the built-in 2-D CCD array, or
- a white stick for blind people using 1-D CCD array (alternatively PSD array) or the handheld device.

The twofold range-inclination tracer consists of the camera with 1-D CCD array or linear PSD element and four laser diodes radiating four intersecting light rays against the wall, as shown in Figure 10. Four light spots are sampled by a simple sampling algorithm, which then evaluates the distance D and the inclination β of the plane against the range-inclination tracer. The sampling of left and right light spot coordinates in subsequent pictures synchronized with sequentially switched light rays enables the recognition of the changing difference and orientation of the light spots and their orientation before and after the crossing point. Control algorithms are derived in the rectangular coordinate system x,y with the origin in the point of mutual intersection of four light rays, as described in Figure 7 using the relationships (4), (5), (6), and (7). The described range-inclination tracer is alternatively used for navigation:

- in single mode, using two crossing light rays with light spots $A(r_1,s_1)$ and $B(r_2,s_2)$, with the evaluation of the distances and inclination of the ambient; or
- in double mode, using four crossing light rays with light spots $A(r_1,s_1)$, $B(r_2,s_2)$, $E(r_3,s_3)$, and $F(r_4,s_4)$, with the evaluation of the distances and inclination of the ambient including the evaluation of the invariant symptoms of the ambient articulation, like simple plain, parallelism, or perpendicularity of two plains.

The navigation system for blind people is based on the cooperation of four laser diodes with focusing optics used to determine light spot in the 1-D CCD array (alternatively PSD array), microprocessor, and tuned pitch and timbre of sound source. The first couple of laser rays with light spots A,B form an angle α, and the second couple of laser rays with light spots E,F form an angle σ; both are symmetric with the axis of the CCD array. Generally the navigation system consists of two independently working range inclination tracers. The configuration of the navigation system enables, by means of purposeful swept hand motion, simple evalu-

Figure 10. Navigation system based on the twofold range-inclination tracer fastened on the handy device



ation of invariant symptoms of the object for the indication of the following:

- simple plane, when the points A,B,E,F belong to one straight line as shown in Figure 11;
- parallel planes, when the points A,E and B,F create two parallel straight lines as shown in Figure 12; or
- two plains forming an angle with pointing up of the right angle, when the points A,E and B,F create two mutually intersecting lines, as shown in Figure 13.

The navigation device can be used in two ways:

1. The single mode using only two light spots A,B or E,F on the way of the range-inclination tracer. This way is effective to assign a safe and acceptable range of the distance and inclination using two-tones. For example, when both light spots are on the pavement, a unique tone is played. The loss of safe inclination is signaled by the dissonance, for example at the positions of the light spots on the pavement-wall. The loss of the front light spot is signaled by the interrupted tone of enhanced intensity.
2. The navigation system using two pairs of the light spots A,B and E,F enables, by means of purposeful swept hand motion, a simple indication of the object articulation. This method is effective to assign various chord combinations (melodies) for different shapes such as simple plane, parallel planes, or two plains forming an angle. Simultaneously the range inclination resonates in the background.

Figure 11. Indication of a simple plain

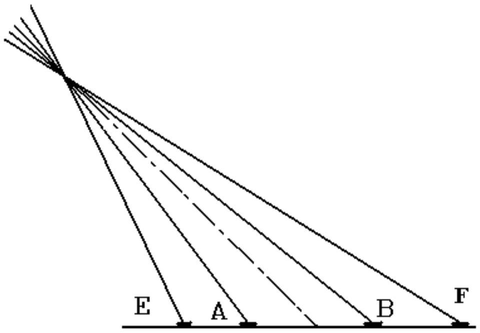


Figure 12. Indication of parallel plains

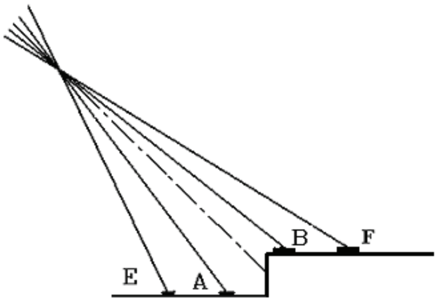
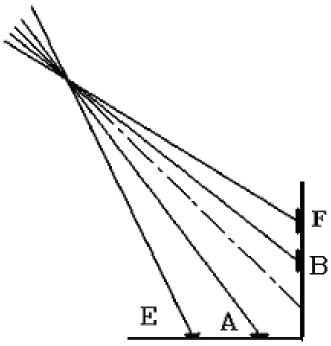


Figure 13. Indication of two plains forming an angle with pointing up of the right angle



Smart configuration of the handy navigator needs simple built-in customizing into a handy phone, handheld device, or white stick for blind people. Video signal from the 1-D or the 2-D CCD array is processed by a simple method at low cost, and in the form of an embedded system, with a single cheap microprocessor in order to decrease the size of opto-electronic devices. The video signal output is preprocessed in the comparator, which is set up on the white level. It causes the signal from every pixel of the light spot to be indicated like an impulse. Every impulse is assigned to the video dot information sequence in the range between 1 and the maximal number of pixels in the CCD array in order to determine its position in the picture coordinate frame. The run of the dot video information is switched for every picture separately by vertical (picture) synchronization impulse. Every couple of light spots A,B and E,F is evaluated in separate pictures. It enables recognition of the varying orientation of the light spot before and after the crossing point of the light rays. An alternative application of the 1-D PSD array is based similarly on the sampling of only one light spot's position.

EXPERIMENTAL HARDWARE FOR COMPOSING OF SIMPLE TASKS

The automatic parking equipment, the range-inclination tracer, the optoelectronic handy navigator, and the wheelchair control by means of the head motion and by means of body motion were implemented by means of following experimental hardware:

- **The Data Translation High-Accuracy, Programmable, Monochrome Frame Grabber Board DT3155 for the PCI Bus:** This is suitable for both image analysis and machine vision applications.
- **The Microprocessor-Controlled Programmable Timer PIKRON ZO-CPU2:** This used for the timing of the light exposure and asynchronous switching of the laser diodes configuration.
- **The Configuration of Miniature Laser Diodes F-LASER 5mW:** This includes focusing optics radiating structured light rays.
- **Digital B/W Video Camera SONY KC-381CG:** This includes a digital signal processor and high-resolution 795Hx596V 1/3" CCD sensor with high sensitivity (0.02lux at F0.75) and interline transfer, digital light level control system for the back-light compensation, with auto-exposure or manual exposure system, aperture correction, and internal or gen-lock/line-lock external transfer.
- **Zoom Lens Computar MLH-10X:** This includes 10x macro zoom, maximal magnification 0.084~0.84x, maximal aperture 1:5.6, maximal image format 6.4×4.8mm (average 8mm), and focus 0.18~0.45m.

The image processing of the light spots from the CCD camera including the control algorithms for the first step of skill-based education was developed on MATLAB. The application of the frame grabber is used only for experiments.

The second step of the education is oriented on the processing of the video signal from the 1-D or the 2-D CCD array for the user's application by a single cheap microprocessor.

The activity of the PSD element is based on the sampling of only one light spot's position. This means that for every light beam, it assigns a separate PSD element. In this way the subsequent sampling of the six-DOF information by means of the PSD element is guaranteed, but causes dynamic distortion dependent on the sampling frequency. Correct activity of the six-DOF sensor based on the CCD element depends on the continuity of sampling. This indicates the verification of every light spot position in regard to responding basic position A,B,C,D,S. The continuous motion of the first light spot can be recognized in various ways, for example by means of:

- different shape of the light spot,
- different color of the light spot,
- different intensity of the light spot,
- dividing of the picture frame into four quadrants for every light spot, and
- minimal distance of preliminary position of the light spot.

Having the identity of the first light spot, the other light spots can be recognized clockwise or counter-clockwise. The third step of the educational applications is oriented toward embedded systems in order to decrease the size of developed optoelectronic devices. The structure of the embedded system is oriented to the use of microprocessors in order to enable flexibility in the proof of various algorithmic modules for the enhancement of the accuracy, like the approximation of the light spot center or for the elimination of the nonlinearity, and on the tasks of the dynamic control using C language with subroutines in assembler.

CONCLUSION

The modular sensory system design presented here enables easy adaptation of various applications of human-machine interfaces for assistive technologies and mobile robotic systems. In general, this modular sensory system concept is appropriate for a low-cost design and in addition enables the understanding of basic problems concerning the interaction between human and mobile robotic systems.

ACKNOWLEDGMENTS

The support from the grant Vyzkumne zamery MSM 7088352102 "Modelovani a rizeni zpracovatelskych procesu prirodnich a syntetickych polymeru" is gratefully acknowledged.

REFERENCES

Akira, I. (2002). An approach to the eye contact through the outstaring game Nirammekko. *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2002)*, Berlin, Germany.

Fukuda, T., Nakashima, M., Arai, F., & Hasegawa, Y. (2002). Generalized facial expression of character face based on deformation model for human-robot communication. *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2002)*, Berlin.

Humusoft. (n.d.). Retrieved from http://www.humusoft.cz

Kawarazaki, N., Hoya, I., Nishihara, K., & Yoshidome, T. (2003). Welfare robot system using hand gesture instructions. *Proceedings of the 8th International Conference on Rehabilitation Robotics (ICORR 2003)*, Daejeon, Korea.

Kim, D.-H., Kim, J.-H., & Chung, M. J. (2001). An eye-gaze tracking system for people with motor disabilities. *Proceedings of the 7th International Conference on Rehabilitation Robotics (ICORR 2001)*, Evry Cedex, France.

Kim, J.-H., Lee, B. R., Kim, D.-H., & Chung, M. J. (2003). Eye-mouse system for people with motor disabilities. *Proceedings of the 8th International Conference on Rehabilitation Robotics (ICORR 2003)*, Daejeon, Korea.

Kvasnica, M. (1999, July). Modular force-torque transducers for rehabilitation robotics. *Proceedings of the IEEE International Conference on Rehabilitation Robotics,* Stanford, CA.

Kvasnica, M. (2001, April). A six-DOF modular sensory system with haptic interface for rehabilitation robotics. *Proceedings of the 7th International Conference on Rehabilitation Robotics (ICORR 2001)*, Paris-Evry, France.

Kvasnica, M. (2001). Algorithm for computing of information about six-DOF motion in 3-D space sampled by 2-D CCD array. *Proceedings of the 7th World Multi-Conference (SCI'2001-ISAS)* (Vol. XV, Industrial Systems, Part II), Orlando, FL.

Kvasnica, M. (2002). Six DOF measurements in robotics, engineering constructions and space control. *Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI'2002-ISAS 2002)* (Ext. Vol. XX), Orlando, FL.

Kvasnica, M. (2003, September). Head joystick and interactive positioning for the wheelchair. *Proceedings of the 1st International Conference on Smart Homes and Health Telematics (ICOST 2003)*, Paris.

Kvasnica, M. (2003). Six-DOF sensory system for interactive positioning and motion control in rehabilitation robotics. *Proceedings of the 8th International Conference on Rehabilitation Robotics (ICORR 2003)*, Daejeon, Korea.

Kvasnica, M. (2003). Six-DOF sensory system for interactive positioning and motion control in rehabilitation robotics. *International Journal of Human-Friendly Welfare Robotic Systems, 4*(3).

Kvasnica, M. (2004, September). Six-DOF force-torque transducer for wheelchair control by means of body motion. *Proceedings of the 2nd International Conference on Smart Homes and Health Telematics (ICOST 2004)*, Singapore.

Kvasnica, M. (2005). Assistive technologies for man-machine interface and applications in education and robotics. *International Journal of Human-Friendly Welfare Robotic Systems, 6*(3). Daejeon, Korea: KAIST Press.

Kvasnica, M., & Vasek, V. (2004, May). Mechatronics on the human-robot interface for assistive technologies and for the six-DOF measurements systems. *Proceedings of the 7th International Symposium on Topical Questions of Teaching Mechatronics,* Rackova Dolina, Slovakia.

Kvasnica, M., & Van der Loos, M. (2000). Six-DOF modular sensory system with haptic interaction for robotics and human-machine interaction. *Proceedings of the World Automation Congress (WAC 2000)*, Maui, HI.

Mathworks. (n.d.) Retrieved form http://www.mathworks.com/

Min, J. W., Lee, K., Lim, S.-C., & Kwon, D.-S. (2003). Human-robot interfaces for wheelchair control with body motion. *Proceedings of the 8th International Conference on Rehabilitation Robotics (ICORR 2003)*, Daejeon, Korea.

Moon, I., Lee, M., Ryu, J., Kim, K., & Mun, M. (2003). Intelligent robotic wheelchair with human-friendly interfaces for disabled and the elderly. *Proceedings of the 8th International Conference on Rehabilitation Robotics (ICORR 2003)*, Daejeon, Korea.

Neovision. (n.d.). Retrieved from http://www.neovision.cz

KEY TERMS

α/2: The angle of the optical axis of the camera with the laser light ray.

B: The inclination of the wall against the wheelchair.

D: The distance between the camera and the wall.

Δh: The difference between the point of the light rays' intersection and the intersection of the optical axis of the camera with the wall.

h: The distance between the camera and the point of the intersection of light rays.

N: The direction of the wheelchair motion.

(r_1,s_1), (r_2,s_2): The coordinates of the light spots A, B.

2s: Presetting control of the angle contained by mutual opposite light rays of the modules Apc and App.

x,y: The coordinates of the position of the wheelchair.

(x_1,y_1), (x_2,y_2): Coordinates of the points E, F.

z: The distance of the wheelchair from the ceiling in the rectangular coordinate frame.

M-Commerce Technology Perceptions on Technology Adoptions

Reychav Iris
Bar-Ilan University, Israel

Ehud Menipaz
Ben-Gurion University, Israel

INTRODUCTION

This article presents a tool for assessing the probability of adopting a new technology or product before it is marketed. Specifically, the research offers managers in firms dealing with mobile electronic commerce a way of measuring perceptions of technology usage as an index for assessing the tendency to adopt a given technology. The article is based on an ongoing study dealing with m-commerce in Israel and internationally. It is centered on creating a research tool for predicting the usage of m-commerce in Israel, based on the PCI model. The suggested model is based on a questionnaire presented to the potential consumer, containing questions linking the consumer's perception of the various aspects of the technological innovation offered, together with his tendency to buy and therefore adopt it. The tool was found to possess high reliability and validity levels. The average score in the questionnaire is used to predict the probability of adoption of the mobile electronic commerce technology. Implications related to m-commerce technology in Israel and worldwide are discussed.

BACKGROUND

The main purpose of this study is to assess the tendency to adopt mobile electronic commerce technologies, prior to actually launching a new product or service based on cellular technology. The focus in the present study is on the general population and not on organizations.

Contrary to products or services sold to end users, the mobile electronic commerce field offers an innovative system of business ties with the client, by utilizing mediating tools such as the cellular device. The focus in the mobile electronic commerce field is on influencing consumers' preferences. This study presents a tool that examines the perceived utilization of technology advances in the field.

The focus in this study on the characteristics of using cellular phones for mobile electronic commerce is based on findings from a wide range of studies dealing with the characteristics of the perception of innovativeness itself.

Rogers (1983) studied thousands of cases of diffusion and managed to define five characteristics of innovativeness affecting its diffusion: relative advantage, compatibility, complexity, visibility, and trialability. While Rogers' characteristics were based on the perception of innovativeness itself, Ajzen and Fishbein (1980, p. 8) claimed that the attitude toward the object is different in essence from the attitude towards a certain behavior related to the object. Innovation penetrates because of accumulating decisions by individuals to adopt it. Therefore, not perceiving the efforts of innovation itself, but the perception of using innovation is the key to its diffusion. In the diffusion studies, the subject of perceptions was treated in relation to innovation itself. Nevertheless, the characteristics of the perceptions of innovation can be redesigned in terms of perceiving the use of the innovation (Moore, 1987). Rewriting the characteristics of *perceptions of innovativeness* into characteristics of the *perceptions of using the innovation* was the basis for the PCI (perceived characteristics of innovating) model, developed by Moore and Benbasat (1991) and used as a tool for studying the adoption of information technologies. The PCI model expands the conceptual framework designed by Rogers, by adding additional characteristics that may influence the decision to adopt a new technology. The tool was presented as reliable and valid.

MOTIVATION AND PURPOSE OF THE STUDY

The motivation for creating a tool for measuring the perceptions regarding cellular phone usage for mobile electronic commerce of the potential adopters of the technology originated from three main factors.

First were findings from previous research, which focused on adoption patterns of Internet and cellular technologies in various countries, in an attempt to present a methodology for analyzing diffusion (Reychav & Menipaz, 2002). Second, while carrying out the above-mentioned study, the researchers realized there was a lack in theoretical background for studying the initial adoption process of innovative technologies

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

such as m-commerce, as well as the understanding of how to successfully assimilate innovative technology. Third, an opportunity presented itself to examine the research model in Israel, a country in which the usage of cellular technology is widespread.

METHOD

The study took place in Israel, where the knowledge constraint towards cellular technology does not exist and apprehension on the part of the general population from adopting unknown technologies due to this constraint is nearly unknown. Outsland (1974, p. 28) suggested that perceptions of innovations by potential adopters of innovative technology might be an effective prediction measure for adoption of the innovation, more than personal factors. Based on this assumption, the current research focused on the university student population, which represents a segment in society that is essentially aware of computer and Internet technologies, and therefore its apprehension from adopting unknown technologies is relatively low. In addition, in Israel the penetration percentage of cellular technology has already reached its full potential, and the interest of the current research is to examine the tendencies to adopt usage of cellular phones for mobile electronic commerce. In order to do so, a research questionnaire was constructed, including 35 items dealing with perceptions regarding the use of mobile electronic commerce technology. Each item in the questionnaire was assessed on three time scales—perceptions of usage in the past, at present, and an estimate of usage perceptions in future.

The questionnaire was distributed amongst students from various departments at Ben Gurion University in the Negev. The distribution included most university departments. A total of 1,300 questionnaires were distributed. They were completed in the presence of the researcher and handed in directly.

CHARACTERISTICS OF THE MODEL

The model is based on the characteristics of the perceptions of innovativeness, which have been identified in previous studies, with a change in wording from "perception of innovativeness" to "perception of the use of innovation," as suggested in the PCI model (Moore & Benbasat 1991). The characteristics are as follows:

- **Relative Advantage:** The extent to which the use of an innovation is perceived as better than the use of its predecessor (based on work by Roger, 1983).
- **Compatibility:** The extent to which the use of an innovation is perceived as being persistent with other existing values, needs, and experiences of the potential adopters (Roger, 1983).
- **Ease of Use:** The extent to which individuals believe that the use of a specific system does not require investment of physical and emotional efforts (Davis, 1986).
- **Results Demonstrability:** The extent to which the results of using an innovation are tangible and presentable (Rogers, 1983, p. 232). Research has shown that merely being exposed to a product can in itself create a positive attitude toward it among individuals (Zajonc & Markus, 1982).
- **Image:** The extent to which the use of an innovation is perceived as improving the individual's status in society.
- **Visibility:** The extent to which the results of the use of an innovation are visible to others. The characteristic "Observability," which was mentioned by Rogers (1983), is presented in the PCI model via two variables (Results Demonstrability and Visibility).
- **Trialability:** The extent to which the use of an innovation can be experienced prior to its adoption.

RESULTS

Out of 1,300 distributed questionnaires, 1,005 were completed correctly (11.49% missing data). The study results point to 55.6% potential users of cellular phones for m-commerce in two years' time, compared to 34.8% two years ago and 38.9% users today. This is indicative of the fact that the questionnaire reflects the target population studied in the current research.

The percentage of explained variance obtained in the model runs is 67.732% in the past, 65.896% at present, and 61.470% in the future.

It is safe to say that the model for this study has been validated and verified. Therefore, we can conclude that in order to assess the probability of actual usage of cellular phone for m-commerce, the model for perceptions of use of cellular phone for m-commerce presented in this study may be used.

Predicting the Probability of Using M-Commerce

After verifying the model, a test was carried out to identify which variables assist in predicting the probability for *using cellular phones for mobile electronic commerce.*

The testing method used was Forward Stepwise Logistical Regression (Hosmer & Lemeshow, 2000), which first brings into the equation variables having the highest level of significance, and then re-calculates the level of significance

Table 1. Logistical regression equation constants in test run on perceptions categories

Variable	*B (past)*	*B (current)*	*B (future)*
Relative Advantage	0.117	0	0
Image	0.096	0	0
Compatibility	0	0.182	0.385
Visibility	0.305	0.273	0.185
Cellular Phone Use	0	0	1.021
Method of Connecting with Internet Provider	0	0	-0.533
Equation Constant	-1.770	-1.501	-2.310

of each of the variables in the equation. For the past timeframe, the results were given after running the first step of the regression, while for the present and future timeframes, the results were given after three steps of the regression.

The results show that for the period of two years ago, the *average perceptions* of potential adopters of the technology is a significant variable ($p<0.01$) and can therefore serve as a tool for predicting the probability of using cellular phones for mobile electronic commerce at that time, and for present and future times as well. For the period of the next two years (future), the variables which were found to be significant are: *estimate of the average perceptions in two years' time* ($p<0.01$), *estimate of cellular phone use,* and *method of connecting to an Internet provider in two years' time* ($p=0.037$).

An assessment of the predictions for using cellular phones for mobile electronic commerce was also done by using each of the categories in the model.

The findings show that for the timeframe of two years ago, the significant perception categories are: *relative advantage* ($p=0.039$), *image* ($p=0.033$), and *visibility* ($p=0.001$). In the present timeframe, the significant categories are: *compatibility* ($p=0.001$) and *visibility* ($p=0.000$), with the *image* category having only a slight significance ($p=0.055$). In the future timeframe, the significant categories are: *compatibility* ($p=0.001$), *visibility* ($p=0.002$), and the additional variables *use of cellular phone* ($p=0.016$) and *method of connecting to an Internet provider* ($p=0.010$).

For each of the significant variables, the logistical regression odds parameters present the probability for prediction of the dependent variable, *use of cellular phone for mobile electronic commerce.* The results show that the categories *relative advantage* and *Image* serve as a prediction tool only for the past timeframe. The *visibility* category serves as a prediction tool for all timeframes, although for each point in the average of the *visibility* category, the probability along the timeframes is lower than 1.357 in the past, 1.314 at present, and 1.203 in future. The *compatibility* category serves as a prediction tool only at present and in the future. Also in the future, additional variables are added which may influence the probability for prediction: *use of cellular phone* (2.775) and *method of connecting to an Internet provider* (0.587). However, the probability presented by the variable *method of connecting to an Internet provider* is lower than 1, therefore reducing the probability for using mobile phones for mobile electronic commerce if the *method of connecting to an Internet provider* is by way of a telephone line.

Given regression coefficients for each one of the test runs performed across the three timeframes, the regression equation can be created so as to predict the probability for *using cellular phone for mobile electronic commerce* for each questionnaire respondent. The results of the regression coefficients are presented in Table 1.

In this case, we can also insert the regression coefficients obtained in each test run in the logistical regression equation, and receive the probability for *using cellular phones for mobile electronic commerce* in each timeframe and for every score received in each of the categories.

For a presentation of the probability to *use cellular phones* according to the average score obtained by every potential adopter of the technology who answers the questionnaire, see Figure 1.

Figure 1 presents a trend in the change in attitudes towards *using cellular phones for mobile electronic commerce* from non-usage to usage. As the *average score of perception of potential adopter* grows from past to present to future, so grows the probability to adopt the technology, from 63%, 79%, and 90% respectively for the maximal score given by the respondent. The variable *method of connecting to an Internet provider,* which was also found to influence the probability of predicting use in the future, heightens the probability of adopting the technology from 90% (when using a telephone line) to 93.485% (when using ADSL).

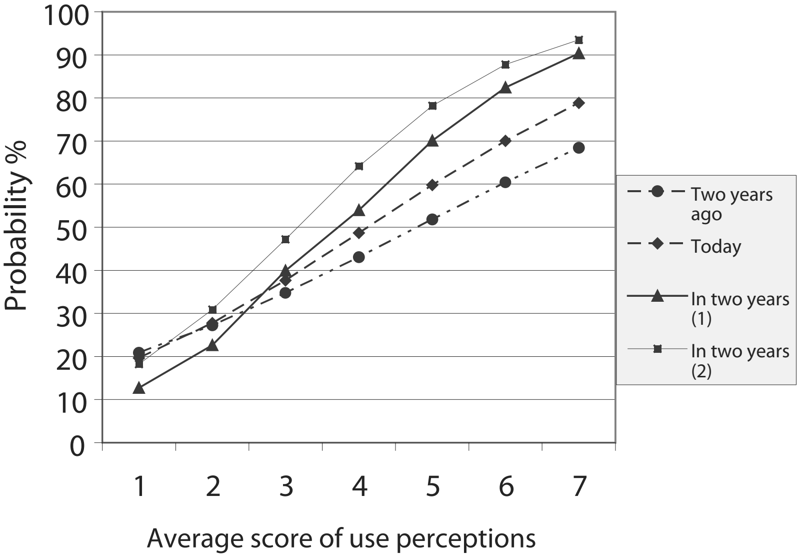
Internet, Cellular, and Credit Card Technology Usage Habits

In addition to the research hypothesis, which dealt with the perceptions model presented in the study, additional research hypotheses were tested, relating to the usage habits of Internet and cellular technologies such as the *use of cellular phones, method of connecting to an Internet provider,* and *use of credit card.*

Cellular Technology

Cellular phone usage rate was 87.1% two years ago. This rate has risen to 95.4% today, and is estimated to change by 0.4%, to reach 95.8% in two years' time.

Figure 1. The probability of using cellular phones for mobile electronic commerce, dependent on the average score of perceptions of using cellular phones for mobile electronic commerce



The results presented here point to the existence of a trend from past to present time only (Sig. (McNemar) p<0.01).

Internet Technology

Regarding the *method of connecting to an Internet provider,* results show that the use of a telephone line as a connecting method has been in decline, from 85.7% two years ago to 53.3% today. This trend is expected to continue with a further decline to 12.8% two years from now. In comparison, the use of ADSL as a connection method has grown by 694.8% over the past two years (from 5.8% two years ago to 46.1% today). It is estimated that this trend will continue in the near future and will reach 84.1% in two years' time (a growth rate of 82.4% compared to the present rate).

The rate of Internet use two years ago was 79.4%. Today it has grown to 97.9%. This rate is expected to grow to 98.4% two years from now.

The results presented here point to the existence of a trend in Internet use from past to present (Sig. (McNemar) p<0.01) and from present to future (Sig. (McNemar) p=0.039). A significant correlation was found, using the chi-square test, between *Internet use* and *cellular phone use* for all timeframes (Sig. (2 tailed) p<0.01). The percentage of cellular phone users found in the study out of the overall Internet users shows a 6% growth in the past two years and an additional 0.8% expected growth in the next two years.

In comparison, the percentage of Internet users out of the overall cellular phone users showed a growth of 16.6% in the past two years, but is expected to decline by 13.1% in the next two years, from a rate of 98.8% users to 85.7% in two years. A further interesting finding is the non-significant correlation between *Internet use* and *use of cellular phones for mobile electronic commerce* for all timeframes.

Credit Card Use

Concerning *credit card use,* the results of the study show that the usage rate of credit cards has risen from 86.1% two years ago to 93.8% today, with an expected rate of 97.3% in two years.

The results presented here point to the existence of a trend in *credit card use* for all timeframes (Sig. (McNemar) p<0.01). The results also point to a significant correlation, by using the chi-square test, between *use of credit cards for Internet use* for all timeframes (Sig. (2 tailed) p<0.01) and the *use of credit cards for cellular phone use* in the past and present times (p<0.01), and also in the future (p=0.015). Similarly to the non-significant correlations obtained between the use of basic technologies for electronic commerce, such as Internet and cellular technologies, the study results also show that the *use of cellular phones for mobile electronic commerce* is not necessarily an indication of the *use of credit cards.*

CONCLUSION

Following are the main conclusions that can be deducted from the study so far:

a. The research tool presented in this study was found to have a high reliability level (Nunnally, 1967), and it may serve as a tool for predicting the probability of *using cellular phones for mobile electronic commerce* in the general population, at an initial adoption environment. The present study attempts to analyze respondents' responses when the decision to adopt a technology is voluntary for the general population. The prediction is deducted from the average score obtained by the respondent in the questionnaire, or from the average score obtained by the respondent in the *visibility* and *compatibility* categories, which were found to be significant in this study.

b. Predicting the *use of cellular phones for mobile electronic commerce* two years from now was found to be influenced not only by perceptions, but by a number of additional factors such as *the use of cellular phone* and *method of connecting to an Internet provider.* In Israel, the *use of cellular phones* has reached its full potential. This finding is also supported by the results of the present study (a 95.4% usage rate today).

c. A significant trend was found in the past, present, and future for transferring from *using a phone line,* to *using ADSL* as the *method of connecting to an Internet provider.* This result points to the fact that consumers who use relatively advanced technologies will also tend to do so regarding new advanced technologies. The logical foundation for this argument is based on the findings obtained from the model of perceptions of using cellular phones for mobile electronic commerce, which related the two categories *compatibility* and *relative advantage* to the same factor in the factor analysis. Hence, ensuring the compatibility of an innovative technology to the needs of potential adopters may also constitute a relative advantage, thus increasing the level of use. The *compatibility* category was also found to have a high level of significance in other diffusion studies (Hurt & Hubbard, 1987). Testing the perceptions of use categories has presented a significant difference between past and present, as well as between present and future times.

d. The findings specified above (c) may also assist in explaining the trend of the innovation of mobile electronic commerce two years ago, which was seen as superior to other available alternatives and was thus identified as having a relative advantage.

e. Additional interesting results found in the study point to a non-significant correlation between *credit card use*, together with the *use of Internet and cellular phone technologies* and *using cellular phones for mobile electronic commerce.* The *use of credit cards* for mobile electronic commerce is perceived as unsafe and requires a high level of functional interaction.

f. The study points to a significant correlation between the *use of Internet* and *the use of cellular technologies.* These two technologies are perceived as complementing each other.

g. The decline in the number of Internet users out of the overall cellular user population, together with the finding that points to a non-significance in the relationship between the *use of Internet* and the *use of cellular phones for mobile electronic commerce,* presents a change towards adopting a more innovative technology, such as the mobile electronic commerce technology. A further verification of this claim can be found in the finding showing a significant relationship between the *method of connecting to an Internet provider* and the *use of cellular phones for mobile electronic commerce.*

h. Similarly to the diffusion theory presented by the Bass model (1969), in which the adopter's population is divided into two groups—innovators and imitators, also in the present research, the population, which tends to use an innovative technology, does not limit itself to using a specific technology, but will continue to do so regarding other advanced technologies such as mobile electronic commerce.

RESEARCH IMPLICATIONS

The research and practical implications of this study include a tool for assessing the probability to adopt an innovative technology or product before it is marketed. Specifically, this study offers managers in companies dealing with mobile electronic commerce to base their assessment of the technologies' adoption chances on the perceptions of use of the technology by potential adopters.

The tool presented in the present study enables predicting the probability of adopting a new technology and modeling the factors influencing its diffusion. From the practical aspect, the study may help decision makers in cellular companies, both as regards infrastructure and service providers, to understand the characteristics of perceptions of using the m-commerce technology in Israel and worldwide, and as a result, to be able to focus their time and efforts on defining appropriate marketing strategies. Performing an analysis from the point of view of the end user, as in absorbing a new information system, is an important factor for assessing the future level of use of a new technology.

This study is one of the first attempts to deal with the diffusion of innovations in the mediating technologies environment. The results of the study open a number of additional research opportunities, such as assessing the usage perceptions of potential adopters of a specific product or service in the m-commerce environment (for instance, designated content services). This future study may serve as an indication to marketers regarding the perceptions of using a product or service while it is in its initial development phase, so that the required adjustments can be made before it is actually launched. An additional interesting research direction would be to examine the tool suggested in the present study on international markets, so as to broaden its scope of generalizations and significance.

REFERENCES

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Bass, F. M. (1969). A new-product growth model for consumer durables. *Management Science, 15*(January), 215-227.

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end user information systems: Theory and result.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, USA.

Hosmer, D. W, & Lemeshow, J. R. S. (2000). *Applied survival analysis regression modeling of time to event data.* New York: Wiley-Interscience.

Hurt, H. T., & Hubbard, R. (1987, May). The systematic measurements of the perceived characteristics of information technologies: Microcomputers as innovations. *Proceedings of the ICA Annual Conference,* Montreal Quebec.

Moore, G. C. (1987). End user computing and office automation: A diffusion of innovation perspective. *INFOR, 25*(3), 214-235.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw Hill.

Olshvsky, R. W. (1980). Time and the rate of adoption of innovation. *Journal of Consumer Research, 6*(March), 425-428.

Ostlund, L. E. (1974). Perceived innovation attributes as predictors of innovativeness. *Journal of Consumer Research, 1*(2), 23-29.

Reychav, I., & Menipaz, E. (2002). M-business diffusion and use: Global perspective. *Proceedings of the 12th Industrial Engineering Conference,* Israel.

Rogers, E. M. (1983). *Diffusion of innovations.* New York: The Free Press.

Zajonc, R. B., & Markus, H. (1982). Affective and cognitive factors in preferences. *Journal of Consumer Research, 9*(September), 123-131.

KEY TERMS

ADSL: Method of connecting to an Internet provider.

Compatibility: The extent in which the use of an innovation is perceived as persistent with other existing values, needs, and experiences of the potential adopters (Roger, 1983).

Ease of Use: The extent in which individuals believe that the use of a specific system does not require investment of physical and emotional efforts (Davis, 1986).

Image: The extent in which the use of an innovation is perceived as improving the individual's status in society.

Perceived Characteristics of Innovating (PCI) Model: Developed by Moore and Benbasat (1991), and is used as a tool for studying the adoption of information technologies.

Relative Advantage: The extent in which the use of an innovation is perceived as better than the use of its predecessor (based on work by Roger, 1983).

Results Demonstrability: The extent in which the results of using an innovation are tangible and presentable (Rogers, 1983, p. 232). Research has shown that merely being exposed to a product can in itself create a positive attitude toward it among individuals (Zajonc & Markus, 1982).

Trialability: The extent in which the use of an innovation can be experienced prior to its adoption.

Visibility: The extent in which the results of the use of an innovation are visible to others. The characteristic "observability," which was mentioned by Rogers (1983), is presented in the PCI model via two variables (results demonstrability and visibility).

M–Learning with Mobile Phones

Simon So
Hong Kong Institute of Education, Hong Kong

INTRODUCTION

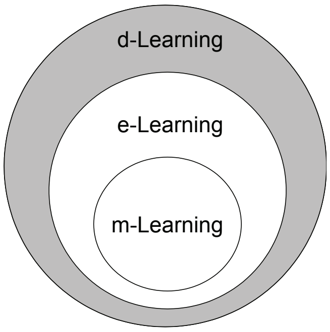
The Internet is a major driver of e-learning advancement and there was an estimate of over 1000 million Internet users in 2004. The ownership of mobile devices is even more astonishing. ITU (2006) reported that 77% of the population in developed countries are mobile subscribers. The emergence of mobile, wireless and satellite technologies is impacting our daily life and our learning. New Internet technologies are being used to support small-screen mobile and wireless devices. In a field marked by such rapid evolution, we cannot assume that the Web as we know it today will remain the primary conduit for Internet-based learning (Bowles, 2004, p.12). Mobile and wireless technologies will play a pivotal role in learning. This new field is commonly known as mobile learning (m-learning).

In this article, the context of m-learning in relation to e-learning and d-learning is presented. Because of the great importance in Web-based technologies to bridge over mobile and wireless technologies, the infrastructure to support m-learning through browser-based technologies is described. This concept represents my own view on the future direction of m-learning. An m-learning experiment, which implemented the concept, is then presented.

BACKGROUND

Many researchers and educators view that m-learning is the descendant of e-learning and originates from d-learning (Wikipedia M-Learning, 2006; Georgiev, Georgieva, & Smrikarov, 2004). The m-learning space is subsumed in the e-learning space and, in turn, in the d-learning space, as shown in Figure 1. This may be true chronologically. D-learning has more than hundred years of evolution starting from the printed media of correspondence (signified by carefully designed and produced materials by specialists to support the absence of instructors and independent study [Charles Wedemeyer] and the industrialization of teaching [Otto Peters]), to mass and broadcast media (marked by the opening of British Open University in 1961 [Daniel, 2001]), and to the telecommunication technologies supporting asynchronous and synchronous learning through teleconferencing, computer mediated communication and online interactive environments for students to create and re-create knowledge individually or collaboratively. In d-learning, the teacher and students are separated quasi-permanently by time, location, or both (Keegan, 2002; ASTD Glossary, 2006). With the advent of computer and communication technologies, e-learning covers a wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms, and digital collaboration (ASTD Glossary, 2006). The delivery of content is through a media-rich and hyperlinked environment utilizing internetworking services. M-learning can be considered as learning taking place where the learner is not at a fixed, predetermined location, or where the dominant technologies are handheld devices such as mobile phones, PDAs and palmtops, or tablet PCs. It can be spontaneous, personal, informal, contextual, portable, ubiquitous and pervasive (Kukulska-Hulme & Traxler, 2005, p. 2).

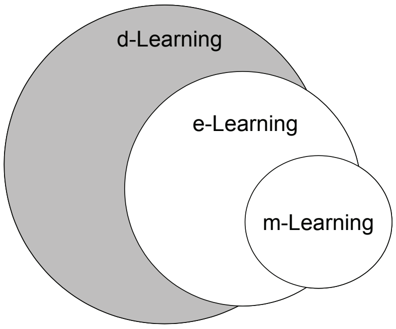
Figure 1. M-learning space as part of e-learning and d-learning spaces



In my view, new concepts in teaching and learning can be generated from m-learning. For example, mobile phones can be used as voting devices for outdoor learning activities or in classrooms without computer supports, as interactive devices in museums, positioning or data logging devices at field trips or in many pedagogical situations. The justification of m-learning being descendent of e-learning and d-learning is rather thin, and Figure 2 is better represented. Furthermore, not everything can be delivered through m-learning. The small form factor, one-finger operation in some cases—slow computational and communication speed, short battery life and limited multimedia capabilities in contrast with computers do not really suit applications requiring heavy reading, high over-the-air communication and a lot of typing or texting.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Overlapping and differential spaces of m-learning, e-learning and d-learning


d-Learning
e-Learning
m-Learning

In summary, m-learning is restricted and expedited by its nature. Different teaching and learning applications require different approaches, whether it is in d-learning, e-learning or m-learning. We must keep in mind their salient characteristics in different teaching and learning contexts, as shown in Table 1.

M-LEARNING INFRASTRUCTURE

In order to support m-learning, mobile devices such as PDAs, mobile phones and tablet PCs, together with servers such as Web servers, streaming servers and database servers on top of applications such as specific adaptation of LMS must be employed (Horton & Horton, 2003; Chen & Kinshuk, 2005). Despite the rapid development in mobile technologies, Figure 3 provides a typical browser-based architecture to support m-learning. It represents a full-scale implementation of any learning system formally. Processing and logic are controlled from the server-side and the mobile devices act as interfaces (Hodges, Bories, & Mandel, 2004, p. 2).

It is also possible that the learning applications are run locally on mobile devices with or without accessing network resources. Applications can be built using Java, such as mobile information device profile (MIDP), C++ on Symbian or native OSs, and Adobe Flash for mobile devices. Feature-rich applications can be implemented to take advantage or avoid limitations of the hardware.

Many researchers believe that, in order to support m-learning, a mobile learning management system (mLMS) is necessary. The logical derivation of mLMS is through the extension of conventional LMS (Trifonova & Ronchetti, 2003; Trifonova, Knapp, Ronchetti, & Gamper, 2004). Direct presentation of materials from computers to mobile devices is likely not legible, aesthetically pleasant, or technically not feasible. Adaptation according to the hardware and device profiles is required. This view is also supported by Goh & Kinshuk (2004). CSS, XSLT and XSL transformation in XML technologies are used to support WML, XHTML and HTML through server pages (Shotsberger & Vetter, 2002). Open standards, including e-learning standards such as SCORM (Fallon & Brown 2003), are the keys for the success of any mLMS.

Table 1. Different teaching and learning contexts

	Salient characteristics
d-learning	• Separation of teachers and learners • Learning normally occurs in a different place from teaching • Formal educational influence and organization
e-learning	• Multimedia-rich • Hypermedia • Independent • Collaborative
m-learning	• Mobile • Portable • Ubiquitous • Pervasive

M-LEARNING WITH MOBILE PHONES

To illustrate the concept discussed above, an m-learning experiment using phone simulators with one of my classes

Figure 3. Browser-based support for m-learning

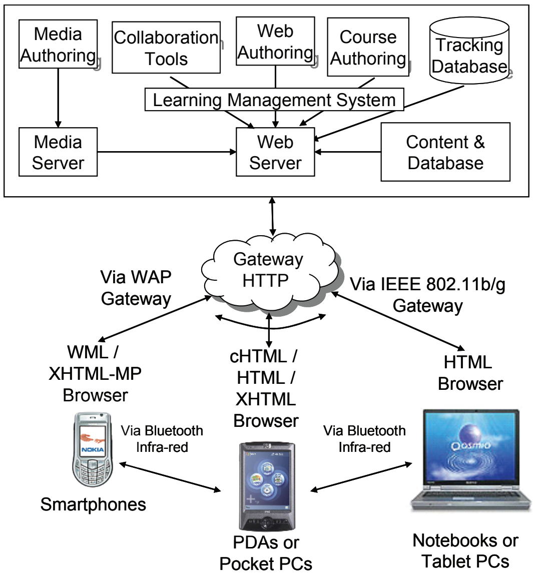

Media Authoring
Collaboration Tools
Web Authoring
Course Authoring
Tracking Database
Learning Management System
Media Server
Web Server
Content & Database
Gateway HTTP
Via WAP Gateway
Via IEEE 802.11b/g Gateway
WML / XHTML-MP Browser
cHTML / HTML / XHTML Browser
HTML Browser
Via Bluetooth Infra-red
Via Bluetooth Infra-red
Smartphones
PDAs or Pocket PCs
Notebooks or Tablet PCs

Figure 4. An m-learning experiment using phone simulators

Figure 5. The system architecture for the m-learning experiment

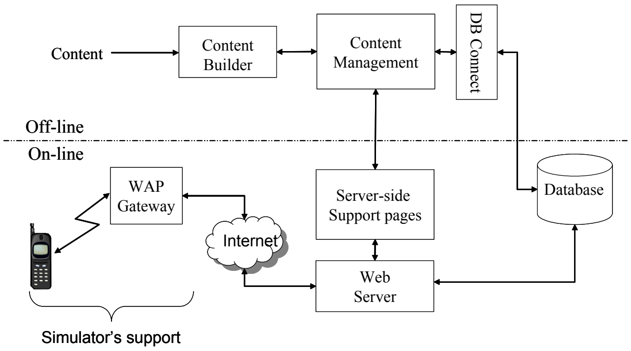


Figure 6. Voting activity

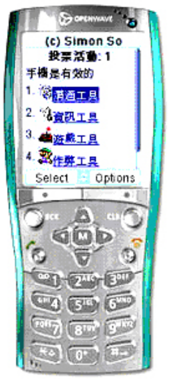
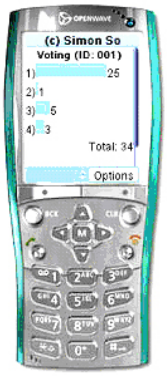


Figure 7. The corresponding voting result



(c) Simon So
Voting (ID: 001)
1) 25
2) 1
3) 5
4) 3
Total: 34

in a computer lab was conducted, as shown in Figure 4. The purpose of the experiment is to find out how my students react to the concept of m-learning. Three activities were developed to address different applications of mobile phones for teaching and learning. Simulators developed to execute in real mobile phones are used for this study (Openwave, 2006). There are three reasons for this. Firstly, the chosen software has been implemented in a number of real phone models. It behaves like a real phone. Secondly, some students may not have mobile phones with advanced features to support WAP 2.0 (Wapforum, 2006) and XTHML-MP, or connect to the mobile service providers with the features turned on. Some students may still have text-based mobile phones! Thirdly, as long as students operate the simulator (e.g., one-finger operation) as the experiment intended, I have a much better controlled environment to answer my research questions.

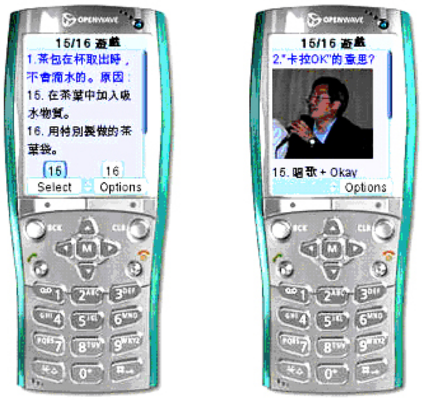
To support this experiment, a WAP gateway connected to a Web server is needed. Figure 5 outlines a practical and partial implementation of the architecture described in the previous section. Apache, PHP and MySQL are chosen as the Web server, server-side programming and database support respectively.

Among the three applications developed for this experiment, the first application is a voting system. Students can cast their votes on their simulators and teachers can interactively check the voting results as illustrated in Figure 6 and Figure 7. Students can use the quick access keys ("1" to "X") on the keypad to cast their votes. This acts as if the voter has a simple voting machine at hand. Teachers can retrieve the voting results from the database onto their handsets as well.

The second application is an interactive game called "15/16" which is a popular game on Hong Kong's television. Instead of two players per game, it was modified that the whole class can participate in each game. Students make their selections and the teacher (or any student) suggests the explanation. Students can change their mind depending on whether they believe the teacher/students or not. Figure 8 illustrates two questions. Teachers can show or refresh the selections at anytime. Figure 9 shows the students' selections for Question #1 in Figure 8.

The third application is a system to administrate tests. Students attempt the questions stored in the database. The overall score can be sent to the students at the end of the test, as shown in Figure 10 and Figure 11. The scores are kept in the database as well.

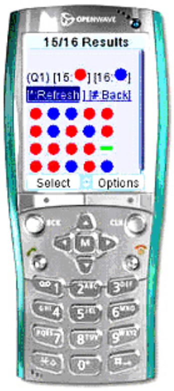
Figure 8. Two questions for the game



15/16 遊戲
1.茶包在杯取出時，不會滴水的。原因：
15. 在茶葉中加入吸水物質。
16. 用特別製做的茶葉袋。
15 16

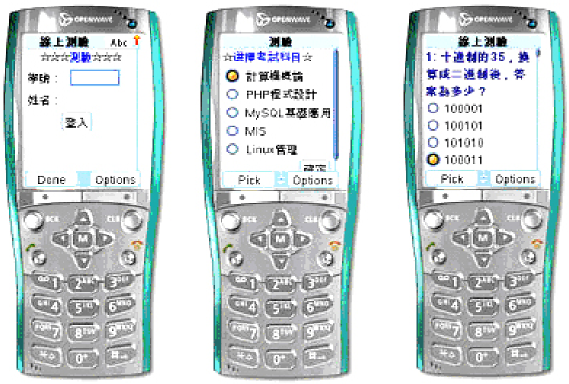
15/16 遊戲
2."卡拉OK"的意思?
15. 唱歌 + Okay

Figure 9. Students' selections



15/16 Results
(Q1) [15:●] [16:●]
[*Refresh] [# Back]

Figure 10. A test on mobile phones



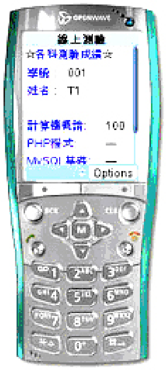
線上測驗
☆☆☆測驗☆☆☆
學號：
姓名：
登入

測驗
☆選擇考試科目☆
● 計算機概論
○ PHP程式設計
○ MySQL基礎應用
○ MIS
○ Linux管理
確定

線上測驗
1: 十進制的35，換算成二進制後，答案為多少？
○ 100001
○ 100101
○ 101010
● 100011

(login) (subject selection) (questions)

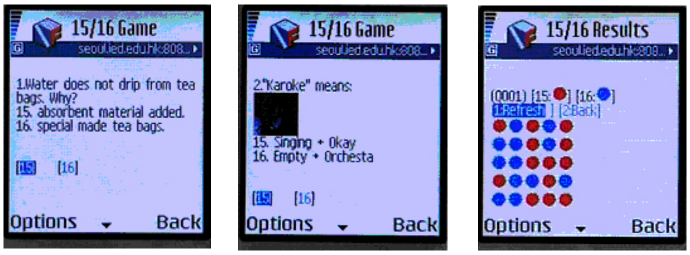
Figure 11. The score



線上測驗
☆各科測驗成績☆
學號 001
姓名： T1
計算機概論: 100
PHP程式 —
MySQL 基礎 —

(student's score)

Figure 12. Nokia's handset implementation corresponding to Figures 8 and 9



15/16 Game
seoul.ied.edu.hk:808...
1.Water does not drip from tea bags. Why?
15. absorbent material added.
16. special made tea bags.
[15] [16]

15/16 Game
seoul.ied.edu.hk:808...
2."Karoke" means:
15. Singing + Okay
16. Empty + Orchesta
[15] [16]

15/16 Results
seoul.ied.edu.hk:808...
(Q001) [15:●] [16:●]
[1.Refresh] [2.Back]

The applications described above are currently being rewritten in English and implemented for Nokia handsets. Figure 12 provides some of the snapshots.

CONCLUSION

M-learning has attracted a lot of research interest recently. It is a fashionable term in education. We can expect a lot of research work in this area will emerge for years to come. It is an exciting field. It also poses a lot of challenges to educators, instructional designs, software engineers and network specialists.

The main concept of m-learning has been highlighted in the article. The browser-based approach to m-learning is presented. It is illustrated by the experiment conducted with my students. This article serves as an example for those researchers to pursue further studies in this direction.

REFERENCES

ASTD Glossary. (2006). *ASTD's source for e-learning*. Retrieved on June 30, 2006, from http://www.learningcircuits.org/glossary.html

Bowles, M. S. (2004). Relearning to e-learn: Strategies for electronic learning and knowledge. Melbourne: Melbourne University Press.

Chen, J., & Kinshuk. (2005). Mobile technologies in educational services. *AACE Journal of Educational Multimedia and Hypermedia*, *14*(1), 89-107

Daniel, J. (2001). The UK Open University: Managing success and leading change in a mega-university. In C. Latchem & D. Hanna (Eds.), *Leadership for 21st Century: Global Perspectives from Educational Innovators*. London: Kogan Page.

Fallon C., & Brown S. (2003). *E-learning standards: A guide to purchasing, developing, and deploying standards-conformant e-learning*. FL.: St. Lucie Press

Georgiev, T., Georgieva, E., & Smrikarov, A. (2004). M-learning: A new stage of e-learning. In *Proceedings of International Conference on Computer Systems and Technologies, CompSysTech'2004*. Retrieved on June 30, 2006, from http://ecet.ecs.ru.acad.bg/cst04/Docs/sIB/428.pdf

Goh, T., & Kinshuk. (2004). Getting ready for mobile learning. In *Proceedings of the 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA2004)* Lugano, Switzerland (pp.56-63).

Hodges, A., Bories, J., & Mandel, R. (2004). Designing applications for 3G mobile devices. In R. Longoria (Ed.), *Designing Software for the Mobile Context: A Practitioner's Guide*. London: Springer.

Horton, W., & Horton, K. (2003). *E-learning tools and technologies: A consumer's guide for trainers, teachers, educators, and instructional designers*. New York: Wiley.

ITU. (2006). Executive summary. In *World Telecommunication/ICT Development Report 2006: Measuring ICT for Social and Economic Development*. Retrieved on June 30, 2006, from http://www.itu.int/ITU-D/ict/publications/wtdr_06/material/WTDR2006_Sum_e.pdf

Keegan, D. (2002*). The future of learning: From eLearning to mLearning*. Retrieved on June 30, 2006, from http://learning.ericsson.net/mlearning2/project_one/book.html

Kukulska-Hulme, A., & Traxler, J. (2005). *Mobile learning: A handbook for educators and trainers*. London: Routledge.

Openwave. (2006). *V7 simulator*. Retrieved on April 15, 2006, from http://www.openwave.com

Shotsberger, P., & Vetter, R. (2002). The handheld Web: How mobile wireless technologies will change Web-based instruction and training. In Allison Rossett (Ed.), *The ASTD E-Learning Handbook: Best Practices, Strategies, and Case Studies for Emerging Field*. New York: McGraw-Hill

Trifonova, A., & Ronchetti, M. (2003). A general architecture for m-learning. In *Proceedings of the Second International Conference on Multimedia and ICTs in Education*. Badajoz, Spain.

Trifonova, A., Knapp, J., Ronchetti, M., & Gamper, J. (2004). Mobile ELDIT: Transition from an e-Learning to an m-Learning. In *Proceedings of the 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA2004)*, Lugano, Switzerland (pp.188-193).

Wapforum. (2006). *WAP 2.0 Standard*. Retrieved on June 30, 2006, from http://www.wapforum.org

Wikipedia M-learning. (2006). *M-learning*. Retrieved on June 30, 2006, from http://en.wikipedia.org/wiki/M-learning

KEY TERMS

Distance Learning (D-Learning): The teacher and students are separated quasi-permanently by time, location, or both. The content can be delivered synchronously or asynchronously.

Electronic Learning (E-Learning): Processes of learning through Web-based learning, online learning, computer-based learning and/or virtual classrooms. The delivery of content is through media-rich and hyperlinked environment utilizing internetworking services.

Learning Management System (LMS): LMS allows the tracking of learner's needs and achievement over periods of time.

Mobile Learning (M-Learning): Learning takes place where the learner is not at a fixed, predetermined location, or where the dominant technologies are handheld devices such as mobile phones, PDAs and palmtops, or tablet PCs.

Wireless Application Protocol 2.0 (WAP 2.0): This is a new version of WAP, which facilitates a cut-down version of XHTML and makes it work in mobile devices.

XHTML Mobile Profile (XHTML-MP): XHTML-MP is a markup language for mobile phones and the like.

Mobile Ad-Hoc Networks

Moh Lim Sim
Multimedia University, Malaysia

Choong Ming Chin
British Telecommunications (Asian Research Center), Malaysia

Chor Min Tan
British Telecommunications (Asian Research Center), Malaysia

INTRODUCTION

Within the coming years, it is inevitable that mobile computing will flourish, evolving toward integrated and converged next generation wireless technology (Webb, 2001), and an important role to play in this technological evolution is mobile ad hoc networks (Liu & Chlamtac, 2004). In short, mobile ad hoc network (MANET) is a self-configuring network that consists of a number of mobile communication nodes that are interconnected by wireless links. The communication nodes are free to move in random manners, which include stationary as a special case. Due to the dynamic movement of the nodes, the ad hoc network topology is normally formed in a decentralized manner and on an ad hoc basis. MANET is in fact a peer-to-peer wireless network that transmits from a client node to another without the use any preexisting network infrastructure-based centralised base stations to coordinate the communications. This type of network is particularly favourable when the information is to be transmitted to other nodes in the locality or the individual communication node has limited radio ranges. The self-configuration, stand alone and quick deployment nature make MANETs suitable for emergency situations like disasters, wars, sporting events, and so forth. Other examples of MANETs with other functionalities include wireless sensor networks and vehicular ad hoc networks.

A wireless sensor network (WSN) is a network formed by a collection of small computers, which are employed in the processing of sensor data (Hać, 2003). These small computers have limited capabilities in terms of the processing and communication power. They usually consist of sensors, a communication device, and a power supply. WSNs find many and varied applications in various fields ranging from industrial monitoring of dangerous environments to agriculture monitoring.

A vehicular ad hoc network (VANET) is a network formed by a collection of vehicles with communications capabilities and also having the potential to support various intelligent transport services. This class of traffic telematics applications ranges from emergency warnings, for example, in the case of accidents, via floating car data gathering and distribution, to more advanced applications, like platooning and co-operative driving (Festag et al., 2004).

In the future, communication devices, communication-capable devices or sensors and home electronic appliances will have the capability to form various MANETs, and interoperate with the global communication networks. These MANETs play an important role in supporting various visions toward the creation of a world of ubiquitous computing where computation is integrated into the environment, rather than having computers that are distinct objects. One of the goals of ubiquitous computing is to enable devices to sense changes in their respective surroundings and to automatically adapt and act on these changes based on user needs and preferences. With ubiquitous computing, people can move around and interact with computers, devices and home appliances more naturally than they currently do.

BACKGROUND

The earliest MANETs were called packet radio networks, and were sponsored by Defense Advanced Projects Agency (DARPA) in the early 1970s (Mobile ad-hoc network, 2006). It is interesting to note that these early packet radio systems predated the Internet, and indeed were part of the motivation of the original Internet Protocol suite. Later DARPA experiments included the Survivable Radio Network (SURAN) project, which took place in the 1980s (Mobile ad-hoc network, 2006). The third wave of academic activity started in the mid-1990s with the advent of inexpensive wireless sensor devices, and Wi-Fi or IEEE 802.11 family of radio cards for personal computers, notebooks and smartphones.

The existing cellular-based broadband access for mobile communications is foreseen to be inefficient due to a number of reasons. Firstly, as the bandwidth required is getting higher approaching hundreds of MHz or tens of GHz range, higher carrier frequency (at least ten times the bandwidth as a rule of thumb) is expected. For a same transmit power level, the wireless channel suffers from greater attenuation as a results of using higher carrier frequency (Etoh, 2005). This calls

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

for the research into the use of multihop communication for the provisioning of broadband access where each hop can support high bandwidth transmission over a short range. Hence MANETs, which is multihop in nature, promise to be one of the most innovative and challenging areas of wireless networking in the future. As mobile technologies are growing at an ever-faster rate, therefore higher reliability and capacity, better coverage and services are required.

The future MANETs will likely evolve along the following directions:

- Different MANETs such as wireless sensor networks, VANETs and infrastructure ad hoc networks are interconnected to form a bigger MANET for better exchange of information.
- The emergence of various radio technologies, such as Bluetooth, UWB, ZigBee, Wi-Fi, WiMAX, and so forth, which are optimized for different functions and with the affordable price of radio cards due to economic of scale, made it practical to install more than one radio card, either of the same type or different, or in a single device. When the communication nodes communicate with each other using more than one radio interface type or channel frequency, it is called multi-radio communication or multi-channel communication.

CHALLENGES OF MANETS

There are a number of technical challenges that need to be addressed in order to ensure good connectivity and quality of service (QoS) for the end-users or client nodes in future MANETs. In the following paragraphs we discuss a few of them.

Dynamic Routing Protocol

Basically, routing protocols with different characteristics may be required for different types of MANETs or under different operating environments. Alternatively, a dynamic routing protocol that can adapt itself to different operating environment is required. For example, conventional routing protocols that have been proven to work fine in MANETs with communication nodes in random movement patterns may not be optimum to support inter-vehicular communications (e.g., VANET) within close proximity that may be moving in cluster form in a specific direction but with micro randomness. Meanwhile, in the case of WSN, conventional MANET protocols such as AODV and DSR may not scale well as the network size increases due to the reservation of large bandwidth for control messages. In addition, the energy limitation of the communication nodes has not been considered (Hać, 2003). The presence of multiple or heterogeneous network interfaces posed a need for an efficient routing mechanism such as multi-radio routing when multiple types of radio technologies are used, or multi-channel routing when different channels of a common radio technology is used. Meanwhile, an appropriate rewarding scheme that helps to accelerate the sharing of resources, which include bandwidth and processing time, among communication nodes in a client ad-hoc network is required.

Network Topology Control

In contrast to wired networks, which typically have fixed network topologies, each communication node in a MANET can change the network topology by adjusting its transmit power or selecting specific nodes to forward its messages, thus controlling its neighbor list. In conjunction with the use of optimum routing algorithms, the challenges of topology control in MANETs are to maintain network connectivity, and optimized network lifetime, throughput, and delay with high scalability, minimum overhead, and high fault tolerance.

In order to achieve high scalability and reduce overhead, formation of sub-groups of nodes among the MANET nodes that perform the routing has been proposed in many algorithms (Perkins et al., 2001; Stojmenovic et al., 2002). In this method, a virtual backbone is formed by using the connected dominating set. The relatively smaller sub-network size helps to reduce the amount of routing information.

A good fault tolerance network may require a fully integrated mesh solution among all communication nodes or through the use of higher transmit power. However, the interference generated may correspondingly degrade the overall network performance in terms of throughput and delay. Thus, there is a trade-off between fault tolerance and capacity performance. In a heterogeneous network environment, the problem becomes worst, as the network topology is governed by the capabilities of diverse types of radio interfaces.

Radio Resource Management

The decentralized nature of MANETs makes it difficult for coordinating the sharing and utilization of radio resources. For each communication node, a large amount of physical parameters are involved, which include the number of radio channels, the type and capability of radio interfaces, the channel conditions (channel quality) that determine the performance of the radio transmission, current communication state (busy or not), and so forth. The collection of communication nodes within a MANET may have different amount of resources for use. Hence, a scheme for the discovery, optimum utilization and scheduling of available resources is required, where further information can be found in Chin et al., 2006.

Power Control and Antenna Beamforming

In a MANET environment, communication nodes tend to interfere with each other due to the decentralized nature of the network formation. This calls for effective methods for interference mitigation. The application of power control in cellular communications has been proven to be able to improve system capacity by reducing unnecessary interference and prolong battery life through reducing the transmit power (Sim et al., 1998; Sim et al., 1999). It can be divided into open loop and closed-loop power control and is widely used in cellular communication systems where base stations will coordinate the power control operations. Adaptation of existing power control algorithms for MANETs is required due to the fact that no centralized node is available for the coordination of power control operations. Without such coordination, all the mobile nodes may greedily transmit at maximum power level for its own sake and hence will cause undue interference to other existing nodes. For a detailed discussion see Olafsson et al. (2005).

Antenna beamforming technique offers a significantly improved solution to reduce interference levels and improve the signal-to-noise ratio through the use of narrower beam in the direction toward the receiving node (Alexiou & Haardt, 2004). With this technology, each mobile node's signal is transmitted and received by the dedicated pair of transmitting and receiving nodes only. The main challenges in using beamforming antennas for MANETs are the cost and physical size. Currently there are efforts on achieving inexpensive beamforming methods, but further works are still required to improve the performance (Liberti & Rappaport, 1999).

Mobility Management

It is anticipated that radio traffics in future wireless networks will be mostly generated by multimedia applications and, hence, next generation networks are expected to provide adequate supports for mobile entertainment with extended geographic coverage. However, multimedia applications often require a dynamic amount of bandwidth and in order to guarantee QoS for such bandwidth-greedy applications when used over a wireless link, current schemes for supporting such services in conventional networks have to be reviewed and new resource management solutions have to be proposed. One of the most critical aspects of guaranteeing QoS support in providing seamless access under dynamic radio conditions is handoff (or handover). A handoff process is either mobile station-triggered or network-triggered, and it involves four successive phases: (1) measurement, (2) handoff initiation, (3) channel assignment, and (4) network connection reconfiguration. Given the prevalent trend toward higher order heterogeneous networks, such a MANET requires advanced schemes to coordinate handoff and reservation of radio resources to ensure the continuity of services to mobile client nodes (Chin et al., 2006).

Security Issues

The MANETs also pose unique challenges in security implementation. This is mostly due to the following properties of some MANETs: resource-constraint mobile nodes, uncontrollable environment and large dynamic network topology. For example, wireless sensor nodes are characterized as severely resource-constraint devices in terms of available power, memory, bandwidth and computational capability. These mobile node-specific factors have set several constraints on the design of security architecture (Sajal et al., 2004). As only a fraction of the total memory may be used by the cryptographic algorithms, the security architecture demands relatively lightweight cryptographic algorithms with a reasonable execution time. The extra overhead required in providing the security service should not substantially degrade the overall efficiency of the MANET.

CONCLUSION

The widespread use of computers and the advancement in embedded system design, microwave techniques and VLSI techniques has stimulated the emergence of various MANETs. The integration of various types of MANET is foreseen in the near future due to the need for pervasive computing. However, the success of such integrated hybrid MANET is very much dependent on the issues discussed herein. Proper solutions are therefore necessary to ensure a successful deployment of MANETs in future wireless environments.

So far the decentralized nature of MANETs has been assumed in various studies. Owing to the complexity of various challenges encountered, there may be a need to look into locally centralized and coordinated MANETs in the future. For example, the gateway node to Internet in a WSN may coordinate the work of topology formation and sensor data aggregation. In addition, the open environments where MANETs will operate in many occasions are uncontrollable and not trustworthy. Hostile circumstances could be envisioned in some situations. Generally, the MANET consists of numerous mobile client nodes organized in a flat or hierarchical structure. Considering the node mobility, authentication and key exchanges must not generate too much overhead messages, since the topology is subject to frequent changes (Schmidt et al., 2005). Additionally, all necessary cryptographic functions and keying material must reside and be executable in the mobile client nodes. Finally, the security architecture needs to be scalable to accommodate a large number of mobile client nodes.

REFERENCES

.

Alexiou, A., & Haardt, M. (2004). Smart antenna technologies for future wireless systems: Trends and challenges. *IEEE Communications Magazine, 42*(9), 90-97.

Chin, C. M., Tan, C. M., & Sim, M. L. (2007). Emerging solutions for optimal utilization of future wireless resources. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce.* Hershey, PA: Idea Group Reference.

Etoh, M. (2005). *Next generation mobile systems 3G and beyond.* West Sussex, UK: John Wiley & Sons.

Festag, A., Fubler, H., Hartenstein, H., Sarma, A., & Schmitz, R. (2004). FleetNet: Bringing car-to-car communication into the real world. In *Proceedings of 11th World Congress on ITS.* Nagoya, Japan.

Hać, A. (2003). *Wireless sensor network designs.* West Sussex, UK: John Wiley & Sons.

Liberti, J. C., & Rappaport, T. S. (1999). *Smart antennas for wireless communications.* Prentice Hall.

Liu, J. J.-N., & Chlamtac, I. (2004). Mobile ad-hoc networking with a view of 4G wireless: Imperatives and challenges. In S. Basagni, M. Conti, S. Giodano, & I. Stojmenovic (Eds.), *Mobile ad hoc networking* (pp. 46). IEEE Press Wiley-Interscience.

Mobile ad-hoc network. (2006). *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Mobile_ad-hoc_ network

Olafsson, S., Freysson, G., Chin, E., & Sim, M.L. (2005, October 6-9). The relevance of adaptive power control for connectivity in de-centralized wireless systems. Paper presented at the the 2005 Networking and Electronic Commerce Research Conference (NAEC 2005), Lake Garda, Italy.

Perkins, C. E., Royer, E. M., Das, S. R., & Marina, M. K. (2001). Performance comparison of two on-demand routing protocols for ad hoc networks. *IEEE Personal Communications, 8*(1), 16-28.

Sajal, K. D., Afrand, A., & Kalyan, B. (2004). Security in wireless mobile and sensor networks. In *Wireless communications systems and networks* (pp. 531-557). Plenum Press.

Schmidt, S., Krahn, H., Fischer, S., & Wätjen, D. (2005). *A security architecture for mobile wireless sensor networks (LNCS)*. Springer Verlag.

Sim, M. L., Gunawan, E., Soh, C. B., & Soong, B. H. (1998). Characteristics of closed loop power control algorithms for a cellular DS/CDMA system. *IEEE Proceedings Communications, 145*(5), 355-362.

Sim, M. L., Gunawan, E., Soong, B. H., & Soh, C. B. (1999). Performance study of close-loop power control algorithms for a cellular CDMA system. *IEEE Transactions on Vehicular Technology, 48*(3), 911-921.

Stojmenovic, I., Seddigh, M., & Zunic, J. (2002). Dominating sets and neighbor elimination-based broadcasting algorithms in wireless networks. IEEE *Transactions on Parallel and Distributed Systems, 13*(1), 14-25.

Webb, W. (2001). The future of wireless communications. Norwood, US: Artech House.

KEY TERMS

AODV: Ad-hoc on demand distance vector.

DARPA: Defense Advanced Projects Agency.

DSR: Dynamic source routing.

MANET: Mobile ad hoc network.

Multihop Communications: A communication mode in which traffic is forwarded to the destination through a number of intermediate communication nodes or routers.

Packet Radio Network: A wireless network that employs packet switching.

Ultra Wide Band (UWB): It is a wireless access technology that uses low power and provides higher speed than Wi-Fi or Bluetooth. It is developed to provide wireless video transmission for home theater systems, cable TV, auto safety and navigation, medical imaging and security surveillance.

VANET: Vehicular ad hoc network.

Very Large Scale Integration (VLSI): It is a microelectronic technology where large-scale systems of transistor-based circuits are integrated into circuits on a single chip.

Wireless Fidelity (Wi-Fi): It is also referred as IEEE802.11. It is a set of standards that set forth the specifications for transmitting data over a wireless network.

WiMAX: World interoperability for microwave access. It is frequently referred to as the IEEE 802.16 wireless broadband standard. It was initially designed to extend local Wi-Fi networks across greater distances, such as a campus, as well as to provide last mile connectivity.

WSN: Wireless sensor network.

ZigBee: A wireless networking technology conforming to the IEEE 802.15.4 standards used for home, building and industrial control and monitoring. It supports the deployment of wireless sensor network. With a slow maximum speed of 250 kbps at 2.4 GHz, ZigBee is slower than Wi-Fi and Bluetooth, but is designed for low power so that batteries can last for months and years. The ZigBee transmission range is short and roughly about 50 meters, but that can vary greatly depending on channel conditions, temperature, humidity and air quality.

Mobile Agent Protection for M-Commerce

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

The introduction of the mobile Internet is probably one of the most significant revolutions of the 20th century. With a simple click, one can connect to almost every corner of the world thousands of kilometers away. This presents a great opportunity for m-commerce. Despite its many advantages over traditional commerce, m-commerce has not taken off successfully. One of the major hindrances is security. The focus of this article is secure transport of mobile agents. A mobile agent is useful for handheld devices like a palmtop or PDA. Such m-commerce devices usually have limited computing power. It would be useful if the users of such devices could send an intelligent, mobile agent to remote machines to carry out complex tasks like product brokering, bargain hunting, or information collection.

When it comes to online transactions, security becomes the primary concern. The Internet was developed without too much security in mind. Information flows from hubs to hubs before it reaches its destination. By simply tapping into wires or hubs, one can easily monitor all traffic transmitted. For example, when Alice uses her VISA credit card to purchase an album from Virtual CD Mall, the information about her card may be stolen if it is not carefully protected. This information may be used maliciously to make other online transactions, thus causing damage to both the card holder and the credit card company.

Besides concerns on security, current m-commerce lacks the intelligence to locate the correct piece of information. The Internet is like the world's most complete library collections unsorted by any means. To make things worse, there is no competent librarian that can help readers locate the book wanted. Existing popular search engines are attempts to provide librarian assistance. However, as the collection of information is huge, none of the librarians are competent enough at the moment.

An intelligent agent is one solution to providing intelligence in m-commerce. But having an agent that is intelligent is insufficient. There are certain tasks that are unrealistic for agents to perform locally, especially those that require a large amount of information. Therefore, it is important to equip intelligent agents with roaming capability.

Unfortunately, with the introduction of roaming capability, more security issues arise. As the agent needs to move among external hosts to perform its tasks, the agent itself becomes a target of attack. The data collected by agents may be modified, the credit carried by agents may be stolen, and the mission statement on the agent may be changed. As a result, transport security is an immediate concern to agent roaming. The SAFE (*s*ecure roaming *a*gent *f*or *e*-commerce) transport protocol is designed to provide a secure roaming mechanism for intelligent agents. Here, both general and roaming-related security concerns are addressed carefully. Furthermore, several protocols are designed to address different requirements. An m-commerce application can choose the protocol that is most suitable based on its need.

BACKGROUND

There has been a lot of research done on the area of intelligent agents. Some literature (Guilfoyle, 1994; Johansen, Marzullo, & Lauvset, 1999) only propose certain features of intelligent agents, some attempt to define a complete agent architecture. Unfortunately, there is no standardization in the various proposals, resulting in vastly different agent systems. Efforts are made to standardize some aspects of agent systems so that different systems can inter-operate with each other. Knowledge representation and exchange is one of the aspects of agent systems for which KQML (Knowledge Query and Manipulation Language; Finin, 1993) is one of the most widely accepted standards. Developed as part of the *Knowledge Sharing Effort,* KQML is designed as a high-level language for runtime exchange of information between heterogeneous systems. Unfortunately, KQML is designed with little security considerations because no security mechanism is built to address common security concerns, not to mention specific security concerns introduced by mobile agents. Agent systems using KQML will have to implement security mechanisms on top of KQML to protect themselves.

While KQML acts as a sufficient standard for agent representation, it does not touch upon the security aspects of agents. In an attempt to equip KQML with built-in security mechanisms, Secret Agent is proposed by Thirunavukkarasu, Finin and Mayfield (1995).

Another prominent transportable agent system is Agent TCL developed at Dartmouth College (Gray, 1997; Kotz et al., 1997). Agent TCL addresses most areas of agent transport by providing a complete suite of solutions. It is probably one of the most complete agent systems under research. Its security mechanism aims at protecting resources and

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the agent itself. In terms of agent protection, the author acknowledges that "it is clear that it is impossible to protect an agent from the machine on which the agent is executing … it is equally clear that it is impossible to protect an agent from a resource that willfully provides false information" (Gray, 1997). As a result, the author "seeks to implement a verification mechanism so that each machine can check whether an agent was modified unexpectedly after it left the home machine" (Gray, 1997). The other areas of security, like non-repudiation, verification, and identification, are not carefully addressed.

Compared with the various agent systems discussed above, SAFE is designed to address the special needs of m-commerce. The other mobile agent systems are either too general or too specific to a particular application. By designing SAFE with m-commerce application concerns in mind, the architecture will be suitable for m-commerce applications. The most important concern is security, as discussed in previous sections. Due to the nature of m-commerce, security becomes a prerequisite for any successful m-commerce application. Other concerns are mobility, efficiency, and interoperability. In addition, the design allows certain flexibility to cater to different application needs.

MAIN FOCUS OF THE ARTICLE

As a prerequisite, each SAFE entity must carry a digital certificate issued by SAFE Certificate Authority, or SCA. The certificate itself is used to establish the identity of a SAFE entity. Because the private key to the certificate has signing capability, this allows the certificate owner to authenticate itself to the SAFE community. An assumption is made that the agent private key can be protected by function hiding (Thomas, 1998). Other techniques were also discussed in the literature (Bem, 2000; Westhoff, 2000), but will not be elaborated in this article.

From the host's viewpoint, an agent is a piece of foreign code that executes locally. In order to prevent a malicious agent from abusing the host resources, the host should monitor the agent's usage of resources (e.g., computing resources, network resources). The agent receptionist will act as the middleman to facilitate and monitor agent communication with the external party.

General Message Format

In SAFE, agent transport is achieved via a series of message exchanges. The format of a general message is as follows:

SAFE Message = Message Content + Timestamp + Sequence Number + MD(Message Content + Timestamp + Sequence Number) + Signature(MD)

The main body of a SAFE message comprises message content, a timestamp, and a sequence number. The message content is defined by individual messages. Here MD stands for the Message Digest function. The first MD is the function applied to Message Content, Timestamp, and Sequence Number to generate a message digest. The second MD in the equation is the application of digital signature to the message digest generated. A timestamp contains the issue and expiry time of the message.

To prevent replay attack, message exchanges between entities during agent transport are labeled according to each transport session. A running sequence number is included in the message body whenever a new message is exchanged.

In order to protect the integrity of the main message body, a message digest is appended to the main message. The formula of the message digest is as follows:

Message Digest = MD5(SHA(message_body) + message_body)

Here SHA (Secure Hash Algorithm) stands for a set of related cryptographic hash functions. The most commonly used function, SHA-1, is employed in a large variety of security applications. The message digest alone is not sufficient to protect the integrity of a SAFE message. A malicious hacker can modify the message body, and recalculate the value of message digest using the same formula and produce a seemingly valid message digest. To ensure the authenticity of the message, a digital signature on the message digest is generated for each SAFE message. In addition to ensuring message integrity, the signature serves as a proof for non-repudiation as well.

If the message content is sensitive, it can be encrypted using a symmetric key algorithm (e.g., Triple DES). The secret key used for encryption will have to be decided at a higher level.

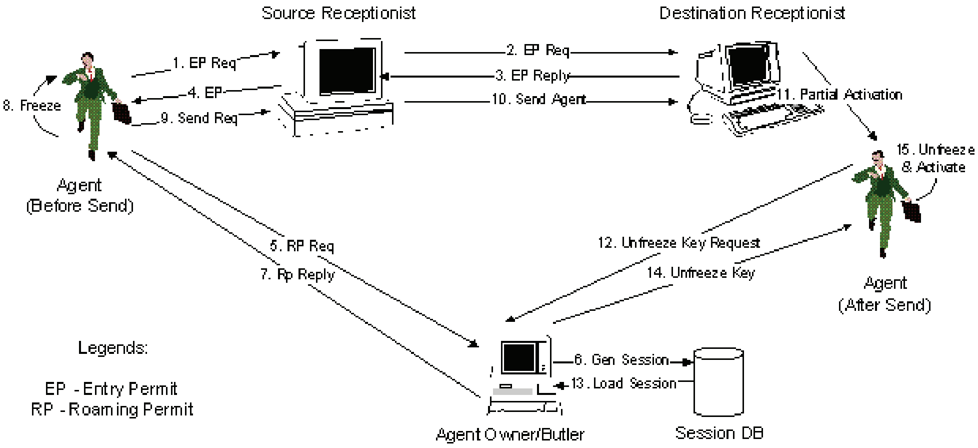
To cater for different application concerns, three transport protocols are proposed: supervised agent transport, unsupervised agent transport, and bootstrap agent transport. These three protocols will be discussed in the following sections in detail.

Supervised Agent Transport

Supervised agent transport is designed for applications that require close supervision of agents. Under this protocol, an agent must request a roaming permit from its owner or butler before roaming. The owner has the option to deny the roaming request and prevent its agent from roaming to undesirable hosts. Without the agent owner playing an active role in the transport protocol, it is difficult to have tight control over agent roaming.

The procedure for supervised agent transport is shown in Figure 1.

Figure 1. Supervised agent transport



Agent Receptionist

Agent receptionists are processes running at every host to facilitate agent transport. If an agent wishes to roam to a host, it should communicate with the agent receptionist at the destination host to complete the transport protocol. Every host will keep a pool of agent receptionists to service incoming agents. Whenever an agent roaming request arrives, an idle agent receptionist from the pool will be activated to entertain the request. In this way, a number of agents can be serviced concurrently.

Request through Source Receptionist for Entry Permit

To initiate supervised agent transport, an agent needs to request for an entry permit from the destination receptionist. Communication between a visiting agent and foreign parties (other agents outside the host, agent owner, etc.) is done using an agent receptionist as a proxy.

Request for Roaming Permit

Once the source receptionist receives the entry permit from the destination receptionist, it simply forwards it to the requesting agent. The next step is for the agent to receive a roaming permit from its owner/butler. The agent sends the entry permit and address of its owner/butler to the source receptionist. Without processing, the source receptionist forwards the entry permit to the address as specified in the agent request.

The agent owner/butler can decide whether the roaming permit should be issued based on its own criteria. If the agent owner/butler decides to issue the roaming permit, it will have to generate a session number, a random challenge, and a freeze/unfreeze key pair. The roaming permit should contain the session number, random challenge, freeze key, timestamp, entry permit, and a signature on all of the above from the agent owner/butler.

In order to verify that the agent has indeed reached the intended destination, a random challenge is generated into the roaming permit. A digital signature on this random challenge is required for the destination to prove its authenticity.

Agent Freeze

With the roaming permit and entry permit, the agent is now able to request for roaming from the source receptionist. In order to protect the agent during its roaming, sensitive function and codes inside the agent body will be frozen. This is achieved using the freeze key in the roaming permit. Even if the agent is intercepted during its transmission, the agent's capability is restricted such that it cannot be run due to the freezing of agent functions. Not much harm can be done to the agent owner/butler. To ensure a smooth roaming operation, the agent's life support systems cannot be frozen. Functions that are critical to the agent's roaming capability must remain functional when the agent is roaming.

Agent Transport

Once frozen, the agent is ready for transmission over the Internet. To activate roaming, the agent sends a request containing the roaming permit to the source receptionist. The source receptionist can verify the validity and authenticity of the roaming permit.

If the agent's roaming permit is valid, the source receptionist will transmit the frozen agent to the destination receptionist as specified in the entry permit. Once the transmission is completed, the source receptionist will terminate the execution of the original agent and make itself available to other incoming agents.

Agent Pre-Activation

When the frozen agent reaches the destination receptionist, it will inspect the agent's roaming permit and the entry permit (contained in the roaming permit) carefully. By doing so, the destination receptionist can establish the following:

1. The agent has been granted permission to enter the destination.
2. The entry permit carried by the agent has not expired.
3. The agent has obtained sufficient authorization from its owner/butler for roaming.
4. The roaming permit carried by the agent has not expired.

If the destination receptionist is satisfied with the agent's credentials, it will activate the agent partially and allow it to continue agent transport process.

Request for Unfreeze Key and Agent Activation

Although the agent has been activated, it is still unable to perform any operation since all sensitive codes/data are frozen. To unfreeze the agent, it has to request for the unfreeze key from its owner/butler. To prove the authenticity of the destination, the destination receptionist is required to sign the random challenge in the roaming permit. The request for unfreeze key contains the session number, the certificate of destination, and the signature on the random challenge.

The direct agent transport process is completed.

Unsupervised Agent Transport

Supervised agent protocol is not a perfect solution to agent transport. Although it provides tight supervision to an agent owner/butler, it has its limitations. Since the agent owner/butler is actively involved in the transport, the protocol inevitably incurs additional overhead and network traffic. This results in lower efficiency of the protocol. This is especially significant when the agent owner/butler is located behind a network with lower bandwidth, or the agent owner/butler is supervising a large number of agents. In order to provide flexibility between security and efficiency, unsupervised agent transport is proposed. The steps involved in unsupervised agent transport are shown in Figure 2.

Request for Entry Permit

In supervised agent transport, session ID and key pair are generated by the agent butler. However, for unsupervised agent transport, these are generated by the destination receptionists because agent butler is no longer online to the agents.

Pre-Roaming Notification

Unlike supervised agent transport, the agent does not need to seek for explicit approval to roam from its owner/butler. Instead, a pre-roaming notification is sent to the agent owner/butler first. It serves to inform the agent owner/butler that the agent has started its roaming. The agent does not need to wait for the owner/butler's reply before roaming.

Agent Freeze

Agent freeze is very close to the same step under supervised agent transport, only that the encryption key is generated by destination instead of agent butler.

Figure 2. Unsupervised agent transport

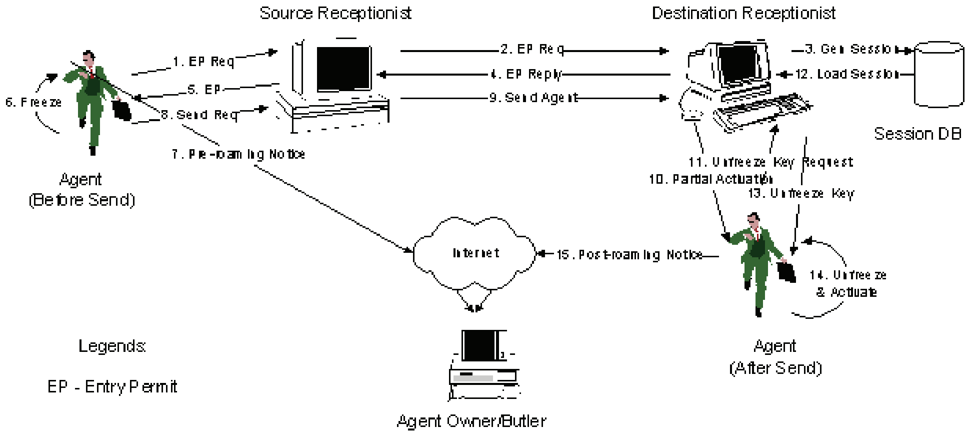
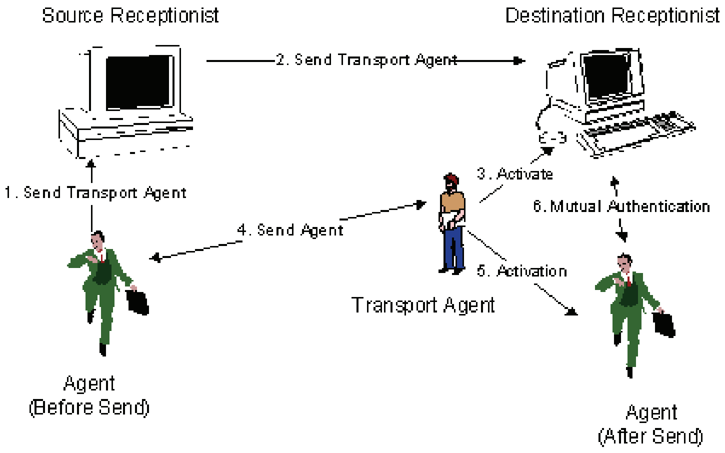

Source Receptionist
Destination Receptionist
1. EP Req
2. EP Req
3. Gen Session
4. EP Reply
5. EP
6. Freeze
7. Pre-roaming Notice
8. Send Req
9. Send Agent
10. Partial Actuation
11. Unfreeze Key Request
12. Load Session
13. Unfreeze Key
14. Unfreeze & Actuate
15. Post-roaming Notice
Session DB
Agent
(Before Send)
Internet
Agent
(After Send)
Legends:
EP - Entry Permit
Agent Owner/Butler

Figure 3. Bootstrap agent transport



Agent Transport

This step is the same as that in supervised agent transport protocol.

Request for Unfreeze Key

The identification and verification processes are the same as compared to supervised agent transport, the exception being that the unfreeze key comes from destination receptionist.

Agent Activation

This step is the same as that in supervised agent transport.

Post-Roaming Notification

Upon full activation, the agent must send a post-roaming notification to its owner/butler. This will inform the agent owner/butler that the agent roaming has been completed successfully. Again, this notification will take place through an indirect channel so that the agent does not need to wait for any reply before continuing with its normal execution.

Bootstrap Agent Transport

Both supervised and unsupervised agent transport make use of a fixed protocol for agent transport. The procedures for agent transport in these two protocols have been clearly defined without much room for variations. It is realized that there exist applications that require a special transport mechanism for their agents. In order to allow this flexibility, SAFER provides a third transport protocol, bootstrap agent transport. Under bootstrap agent transport, agent transport is completed in two phases. Bootstrap agent transport is illustrated in Figure 3.

In the first phase, the transport agent is sent to the destination receptionist using either supervised or unsupervised agent transport with some modifications. The original supervised and unsupervised agent transport requires agent authentication and destination authentication to make sure that the right agent reaches the right destination. Under bootstrap agent transport, the transmission of transport agent does not require both agent authentication and destination authentication.

Once the transport agent reaches the destination, it starts execution in a restricted environment. It is not given the full privilege as a normal agent because it has yet to authenticate itself to the destination. This is to prevent the transport agent from hacking attempts to local host. Under the restricted environment, the transport agent is not allowed to interact with local host services. It is only allowed to communicate with its parent until the parent reaches destination. SAFER allows individual transport agents be customized to use any secure protocol for parent agent transmission.

When the parent agent reaches the destination, it can continue the handshake with the destination receptionist and perform mutual authentication directly. The authentication scheme is similar to that in supervised/unsupervised agent transport.

FUTURE TRENDS

As an evolving effort to deliver a more complete architecture for agents, SAFER (secure agent fabrication, evolution, and roaming) architecture is being proposed to extend the SAFE architecture. In SAFER, agents not only have roaming capability, but can make electronic payments and can evolve to perform better.

CONCLUSION

SAFE is designed as a secure agent transport protocol for m-commerce. The foundation of SAFE is the agent transport protocol, which provides intelligent agents with roaming capability without compromising security. General security concerns as well as security concerns raised by agent transport have been carefully addressed. The design of the protocol also takes into consideration differing concerns for different applications. Instead of standardizing on one transport protocol, three different transport protocols are designed, catering to various needs. Based on the level of control desired, one can choose between supervised agent transport and unsupervised agent transport. For applications that require a high level of security during agent roaming, bootstrap agent transport is provided so that individual applications can customize their transport protocols. The prototype of SAFE agent transport protocol has been developed and tested.

REFERENCES

Bem, E. Z. (2000). Protecting mobile agents in a hostile environment. *Proceedings of the ICSC Symposia on Intelligent Systems and Applications (ISA 2000)*.

Corley, S. (1995). The application of intelligent and mobile agents to network and service management. *Proceedings of the 5th International Conference on Intelligence in Services and Networks,* Antwerp, Belgium.

Finin, T. (1994). *KQML—A language protocol for knowledge and information exchange.* Technical Report CS-94-02, University of Maryland, USA.

Finin, T., & Weber, J. (1993). *Draft specification of the KQML agent communication language.* Retrieved from http://www.cs.umbc.edu/kqml/kqmlspec/spec.html

Gray, R. (1997). *Agent TCL: A flexible and secure mobile-agent system.* PhD Thesis, Department of Computer Science, Dartmouth College, USA.

Guan, S. U., & Yang, Y. (1999). SAFE: Secure-roaming agent for e-commerce. *Proceedings of CIE'99,* Melbourne, Australia (pp. 33-37).

Guilfoyle, C. (1994). *Intelligent agents: The new revolution in software.* London: OVUM.

Johansen, D., Marzullo, K., & Lauvset, K.J. (1999). An approach towards an agent computing environment. *Proceedings of the ICDCS'99 Workshop on Middleware.*

Kotz, D., Gray, R., Nog, S., Rus, D., Chawla, S., & Cybenko, C. (1997). Agent TCL: Targeting the needs of mobile computers. *IEEE Internet Computing, 1*(4), 58-67.

Odubiyi, J. B., Kocur, D. J., Weinstein, S. M., Wakim, N., Srivastava, S., Gokey, C., & Graham, J. (1997). SAIRE—A Scalable Agent-Based Information Retrieval Engine. *Proceedings of the Autonomous Agents 97 Conference* (pp. 292-299), Marina Del Rey, CA.

Rus, D., Gray, R., & Kotz, D. (1996). Autonomous and adaptive agents that gather information. *Proceedings of the AAAI '96 International Workshop on Intelligent Adaptive Agents.*

Rus, D., Gray, R., & Kotz, D. (1997). Transportable information agents. In M. Huhns & M. Singh (Eds.), *Readings in agents.* San Francisco: Morgan Kaufmann.

Sander, T., & Tschundin, C.F. (1998). Protecting mobile agents against malicious hosts. *Mobile Agents and Security, LNCS 1419,* 44-60.

Schneider, F. B. (1997). Towards fault-tolerant and secure agentry. *Proceedings of the 11th International Workshop on Distributed Algorithms,* Saarbrücken, Germany.

Schneier, B. (1996). *Applied cryptography: Protocols, algorithms, and source code in C* (2nd ed.). New York: John Wiley & Sons.

Schoonderwoerd, R., Holland, O., & Bruten, J. (1997). Ant-like agents for load balancing in telecommunications networks. *Proceedings of the 1997 1st International Conference on Autonomous Agents* (pp. 209-216). Marina Del Rey, CA.

Thirunavukkarasu, C., Finin, T., & Mayfield, J. (1995). Secret agents—a security architecture for the KQML agent communication language. *Proceedings of the CIKM'95 Intelligent Information Agents Workshop,* Baltimore, MD.

Westhoff, D. (2000). On securing a mobile agent's binary code. *Proceedings of the ICSC Symposia on Intelligent Systems and Applications (ISA 2000)*.

White, D. E. (1998). *A comparison of mobile agent migration mechanisms.* Senior Honors Thesis, Dartmouth College, USA.

KEY TERMS

Agent: A piece of software that acts to accomplish tasks on behalf of its user.

Digital Certificate: Certificate that uses a digital signature to bind together a public key with an identity—information such as the name of a person or an organization, his or her address, and so forth. The certificate can be used to verify that a public key belongs to an individual

Electronic Commerce (E-Commerce): Consists primarily of the distributing, buying, selling, marketing, and servicing of products or services over electronic systems such as the Internet and other computer networks.

Encryption: The art of protecting information by transforming (*encrypting*) it into an unreadable format, called cipher text. Only those who possess a secret *key* can decipher (or *decrypt*) the message into plain text.

Flexibility: The ease with which a system or component can be modified for use in applications or environments other than those for which it was specifically designed.

Mobile Commerce (M-Commerce): Electronic commerce made through mobile devices.

Protocol: A convention or standard that controls or enables the connection, communication, and data transfer between two computing endpoints. Protocols may be implemented by hardware, software, or a combination of the two. At the lowest level, a protocol defines a hardware connection

Security: The effort to create a secure computing platform, designed so that agents (users or programs) can only perform actions that have been allowed.

Mobile Agent-Based Discovery System

Rajeev R. Raje
Indiana University Purdue University Indianapolis, USA

Jayasree Gandhamaneni
Indiana University Purdue University Indianapolis, USA

Andrew M. Olson
Indiana University Purdue University Indianapolis, USA

Barrett R. Bryant
University of Alabama at Birmingham, USA

INTRODUCTION

For reasons of economy and scalability, many of the current distributed computing systems (DCSs) are realized as an integration of prefabricated and deployed components offering specific services. A critical task that the assembler of such a system needs to address is to locate and select appropriate components scattered over a network. This requires solving many research challenges. These include: (a) deployment of components and their specifications, (b) efficient searching for and gathering of appropriate specifications, (c) representation of queries, and (d) semantics of matching between queries and specifications. UniFrame (Raje, Auguston, Bryant, Olson, & Burt, 2001) is a framework that allows the seamless discovery and integration of such distributed software components. It addresses three key research issues: (1) architecture-based interoperability, (2) distributed discovery of resources, and (3) quality validation. This article presents a mobile-agent-based discovery service, which is one of the alternatives developed under research issue (2).

BACKGROUND

There have been many attempts at creating discovery services. This section reviews only a few prominent ones for the sake of brevity.

Jini (Waldo, 1999) is based on the underlying Java Remote Method Invocation infrastructure (Sun Microsystems, 1994), and thus provides a simplified interoperability. Services register themselves in Lookup Registries, which clients search to download their required services. The matching used in Jini is based on attribute comparisons.

The model used in the Ninja secure service discovery service (SSDS) (Czerwinski, Zhao, Hodes, Joseph, & Katz, 1999) to locate an appropriate service for a request is based on the concept of advertisement. SSDS tracks services in a network and allows authenticated users to locate them through expressive queries. It uses XML to describe the services and to allow complex queries. It supports the possibility of describing various attributes, such as the quality of service (QoS) parameters and associated costs, which are used in the matching process.

CORBA® (Common Object Request Broker Architecture) includes the Trader service (OMG, 2000), which uses a standardized Interface Definition Language to describe service interfaces. These interfaces provide the basis on which lookup and client invocations take place. The trader provides a simple attribute matching.

The aim of Agora (Seacord, Hissam, & Wallnau, 1998) is to provide an automatically generated, indexed, worldwide database of software products classified by their types. Agora combines introspection with Web search engines to reduce the costs of seeking components in the software marketplace. The query terms used for finding components are compared against the index collected by the search engines. The result is inspected by the user so the search can be broadened or refined based on the number and quality of matches.

Universal description, discovery, and integration (UDDI) (OASIS Consortium, 2000) defines a set of services supporting the description and discovery of businesses, organizations, and other Web service providers, as well as the Web services they provide. It utilizes Web Services Description Language (WSDL) for describing the capabilities of the services. UDDI provides a simple textual matching process by comparing each search term with various fields in a service's description.

Web services peer-to-peer discovery service (Banaei-Kashani, Chen, & Shahabi, 2004) is a decentralized discovery service with a matching capability that extends up to the semantic level. It is used to locate Web services that are geographically dispersed across a network. It uses keywords and semantically annotated WSDL to describe Web service interfaces. Each entity, called a "Servent," in this environment

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

serves as both client and server. When a Servent receives a query for a Web service that is not available locally, it shifts its capacity from server to client and queries the network for that specific request. For discovery purposes, a Servent formulates a query encapsulated in a simple object access protocol (SOAP) message (W3C, 2004) and propagates it over the network based on a probabilistic flooding dissemination mechanism.

SLP (Guttman, 1999) provides hosts with access to information about the existence, location, and configuration of networked services. In this framework, user agents model client applications, service agents advertise services, and directory agents cache service information. A user agent can issue service requests to specify the requirements of the client application. It can transmit a request to service agents or a directory SLP. The SLP supports matching only at the syntactical level.

The monitoring and discovery service (MDS) (Globus Alliance, n.d.; Kandagatla, 2003), a part of the Globus Toolkit, is used for discovering computational resources deployed in a Grid environment. The resources are described using a standard schema made up of keywords and can be discovered using specific characteristics. The MDS is made up of two components: the Grid information resource service (GRIS) and the Grid index information service (GIIS). The GRIS runs on resources deployed on the Grid and is an information provider framework for specific information sources. A GIIS is a user-accessible directory server at a higher level that accepts information from child GIIS and GRIS instances and aggregates it for use at a higher level.

THE MOBILE UNIFRAME RESOURCE DISCOVERY SERVICE

A majority of the previously mentioned approaches for discovering service-providing components use relatively simple schemes for describing and matching services against a request. Also, none of these alternatives uses mobile agents in the discovery process. This section provides details about the Mobile UniFrame Resource Discovery Service (MURDS), which has a hierarchical architecture and uses mobile agents to discover services deployed over a network. MURDS is an enhancement of the UniFrame Resource Discovery Service (URDS) (Siram, 2002).

URDS is a hierarchical discovery service that supports the proactive discovery of component specifications, resolves technological heterogeneity, and allows multi-level matching. URDS is one of the entities in the UniFrame approach for developing DCS from heterogeneous, distributed software components. The core concept behind UniFrame is the unified meta-component model (UMM). UMM, as described in Raje (2000), consists of: (a) component, (b) services, and (c) infrastructure.

A component in UniFrame is developed by following a specification in a standardized knowledgebase (KB) (Raje et al., 2001) and implemented in any distributed-component technology. In addition to the implementation of a component, its developer must create, following the specification format described in the KB, a comprehensive specification for it. This is the UMM specification for that component. This, as indicated in Olson, Raje, Bryant, Burt, and Auguston (2005), consists of multiple levels—syntax, semantics, synchronization, and quality of service. Such a complex specification supports multi-level matching while seeking components for a specific query.

Each component in UniFrame offers a service whose UMM description provides its specification. In addition to the functionality of the service, UniFrame emphasizes the service's QoS aspect. Each component indicates its QoS using parameters described in the UniFrame QoS catalog (Brahnmath, 2002).

The infrastructure part of the UMM defines the necessary computational fabric on which components can be deployed and their specifications can be advertised. This allows a proactive discovery of the components for specific queries. URDS provides this infrastructure in UniFrame.

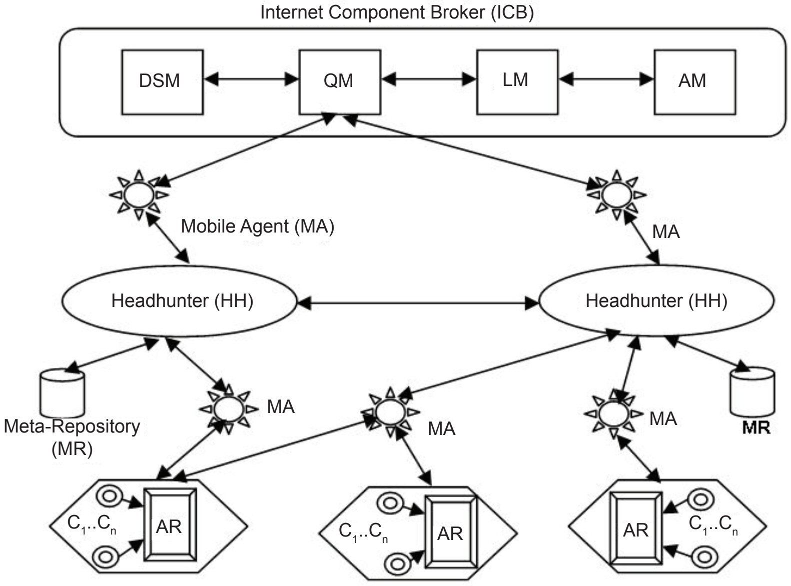
Incorporating the use of mobile agents in the discovery process into URDS creates MURDS. Its architecture, shown in Figure 1, comprises the following entities: (a) Internet Component Broker (ICB), (b) Headhunters (HHs), (c) Meta-Repositories (MR), (d) Active Registries (AR), (e) Components ($C_1..C_n$), and (f) mobile agents (MA). Figure 1 also depicts their interactions. Gandhamaneni (2004) discusses these entities in detail. A brief description follows below.

Internet Component Broker (ICB)

The ICB is a collection of the following entities: q*uery manager (QM), domain security manager (DSM), link manager (LM),* and *adapter manager (AM).* The ICB is a component broker that is pervasive in an interconnected environment. It is expected that there will be a number of ICBs deployed at well-known locations hosted by organizations supporting the UniFrame paradigm for developing distributed systems. The functions of the entities that make up the ICB are as follows:

- **Domain Security Manager (DSM):** The DSM serves as an authorization controller that handles the generation and distribution of the secret keys needed for communication between various constituents of MURDS. It enforces group memberships and performs access control checks on HHs on behalf of the ARs. For performing access control checks, the DSM has a repository of valid users (i.e., HHs, MAs acting on behalf of HHs, and the ARs) and the policies that

Figure 1. MURDS architecture



regulate the associations among them (i.e., the ARs and the HHs).

- **Query Manager (QM):** The QM is responsible for propagating the component selection queries it receives from a user to 'appropriate' HHs. The QM accomplishes this by sending a mobile agent on its behalf to select a list of service provider components that match the search criteria in the query. The current MURDS prototype bases *appropriateness* on the application domain specified in the search requirements. However, more complex schemes, say that use past performance, can be employed to decide to which HHs to send the queries.
- **Link Manager (LM):** The LM establishes links between different ICBs to form a federation of ICBs. Such a federated approach provides a much larger search space for discovering components. An ICB administrator configures the LM with the location information of other ICBs with which links are to be established. Then the QM and the LM can propagate the queries to the other linked ICBs as necessary.
- **Adapter Manager (AM):** The AM acts as a lookup service for clients needing adapter components. These adapter components assist in resolving technological heterogeneity that may occur between two communicating components.

Headhunters (HHs)

The HHs are responsible for proactively detecting, with the help of mobile agents, the presence of components offering services and registering their functionalities in their respective meta-repositories. After receiving a query from a QM, an HH searches its meta-repository and returns a list of components that match the query. The component selection performed by the HH can be based on the concepts of multi-level matching, which aims to match the query requirements with different levels present in the UMM specification. Thus, the matching that the HHs perform (and hence, the MURDS) is much more comprehensive than simple attribute-based matching—a scenario utilized by various discovery services described earlier.

Active Registry (AR)

The AR is an enhancement of the native registry/lookup mechanism that is present in a distributed computing model. For example, in the case of Java-RMI, the AR is a modification of the built-in naming service so that it can listen to broadcasts from HHs and permit mobile agents to access the information about its registered components. ARs have introspection capabilities so that they can provide the specifications of the components registered with them.

Meta-Repository (MR)

The MR is a database belonging to an HH for storing the UMM specifications of the various components the HH finds. Each HH continually attempts to discover new components available on the network to populate its MR.

Mobile Agents (MA)

Mobile agents act as proxies for headhunters in discovering components and for query managers in propagating queries. The MAs that a headhunter sends carry and present that headhunter's credentials to the active registries and seek components from them to be sent back to the meta-repository of that headhunter. The MAs that a query manager sends carry the incoming query to a set of headhunters (or ICBs via the link manager). The addition of the MAs distinguishes the MURDS from the URDS.

Components (C_1…C_n)

The components offering services, which are deployed on the network, may be implemented in different distributed component models. Each of these (and hence, its service) registers itself with its corresponding AR by providing its type name and associated UMM specification.

MURDS' Method of Operation

Developers of a DCS are users, or clients, of the MURDS system. Their goal is to obtain components that match certain functional and non-functional requirements for use in their development process. MURDS receives an incoming query from a user via its QM. Once the QM receives the query, it determines a subset of HHs to which to propagate the query. The MURDS prototype described later selects this subset randomly. After the QM identifies the subset, it sends MAs to the HHs in it. On receiving the query, each HH checks its MR for matching components. It returns any present via the MA to the QM. Also, periodically an HH sends MAs to a set of ARs in order to discover components newly registered with these ARs. An AR, after acknowledging an MA from an HH, then decides which specific type of access, if any, to grant the MA based on the HH's credentials the MA carries. The type of access provided depends upon the policy decisions enforced by the owner of that AR. These policy decisions can be dynamic, and hence the components revealed by an AR to a MA can vary over time. After gathering the desired component specifications from an AR, an MA may choose to send them back to the HH immediately, or continue to other ARs and send a consolidated collection at the end of its journey.

DESIGN, IMPLEMENTATION, AND EXPERIMENTATION

The design of a prototype of MURDS follows a multi-tier architecture, typical of distributed applications, as Figure 2 exhibits. The architecture consists of three tiers, namely the client, middle, and database. The prototype's *client tier* supports application clients. The *middle tier* supports client services (i.e., DSM, HH, QM, and AR) and Grasshopper™ version 2.2.4-enabled mobile agents (Magedanz, Bäumer, & Choy, n.d.). The Java 2 Platform, Enterprise Edition (J2EE)™ version 1.4.2 (Sun Microsystems, 2001) software environment implements various components of the MURDS prototype. The core architectural components (DSM, QM, HH, and AR) are Java-RMI-based services. The *database*

Figure 2. Design of a prototype of MURDS

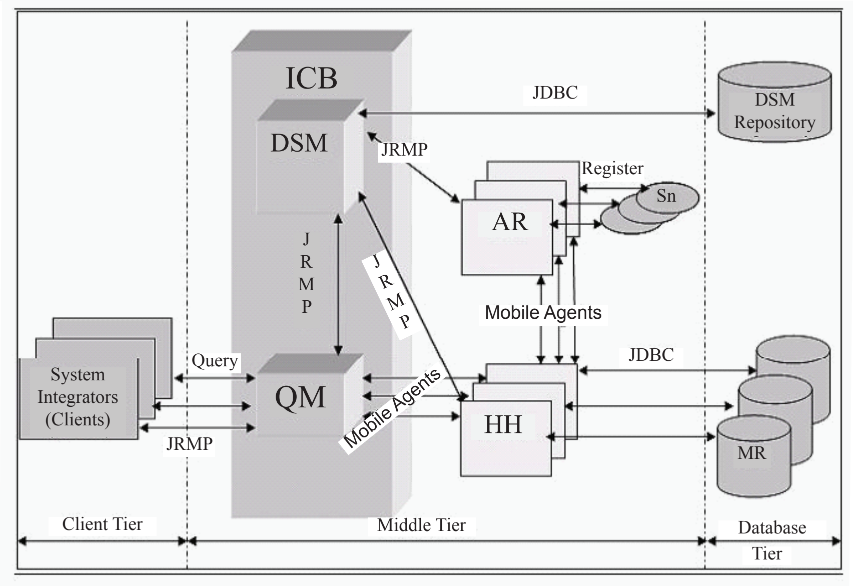
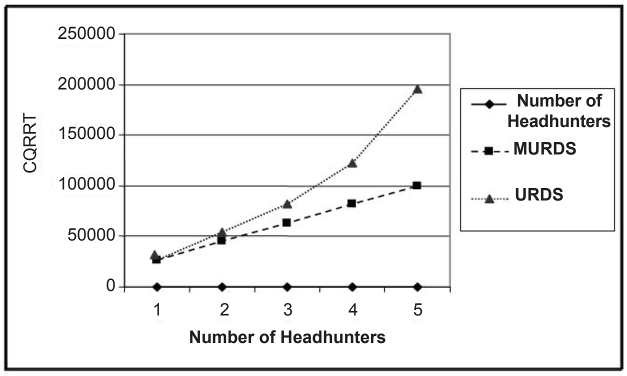


Figure 3. Effect of the number of HHs on CQRRT



tier supports access to the repositories by means of standard APIs. The repositories (DSM-Repository and Meta-Repositories) are Oracle ™ v. 9.2 databases.

Various experiments were conducted to evaluate the prototype of MURDS. Due to space constraints, this article describes only one. Gandhamaneni (2004) presents the details about the others. An important metric these studies used to judge the performance of the prototype was client query result retrieval time (CQRRT). CQRRT is defined as the time taken from the instant a client issues a query to MURDS until the instant the results return back to the client. The experiments investigated the impact of various parameters, such as the number of components, of HHs, and of ARs on the CQRRT.

The result of one such experiment carried out to investigate the effect of increasing the number of HHs on the CQRRT appears in Figure 3. It also compares the results from the MURDS prototype (mobile agents) with the results from the URDS prototype (stationary agents). The configuration used for this experiment consisted of one QM, one client, and seven ARs. As Figure 3 reveals, increasing the number of headhunters increases the CQRRT for both the mobile-agent-based and the stationary-agent-based component selection processes. The reason is that the agents in both processes contact one headhunter at a time to retrieve component information from that HH. Each headhunter consumes time to retrieve components from its local meta-repository and to return them to the contacting agents. As the number of headhunters specified in the search list increases, the amount of time that it takes to return results to the query manager increases, which increases the CQRRT.

Figure 3 indicates, in addition, that the stationary-agent-based component selection process consumes more time to retrieve results from a set of headhunters than the mobile-based-component selection process. This is attributable to the synchronous communication mode of the former vs. the asynchronous mode of the latter, because these modes are the only differences between the two processes.

FUTURE TRENDS

Many extensions to the architecture of MURDS are possible. These could involve comprehensive study of its scalability and classifying the HHs and ARs into categories for deciding the mobile agents' itineraries, or using multi-level matching criteria in selecting appropriate services. The ARs could provide differentiated service and cost frameworks to the agents. An incorporation of these features into MURDS would provide a more comprehensive discovery service that uses mobile agents to search for components deployed over a network. In this way, it would anticipate demands that discovery services will face in the near future.

CONCLUSION

During the assembly of a distributed computing system, the importance of discovering the most appropriate components from those available over a network cannot be overstated. Such discovery, due to its inherent complexity, presents many interesting challenges. This article has briefly described one possible approach, which uses the concepts of UniFrame and mobile agents to identify appropriate components that are scattered over a network. The prototype's results have demonstrated the feasibility of this approach, and further investigations are currently underway as a part of ongoing UniFrame research.

REFERENCES

Banaei-Kashani, F., Chen, C., & Shahabi, C. (2004). *WSPDS: Web services peer-to-peer discovery service.* Retrieved March 22, 2006, from http://infolab.usc.edu/DocsDemos/isws2004_WSPDS.pdf

Brahnmath, G. (2002). *The UniFrame quality of service framework.* Unpublished MS thesis, Department of Computer and Information Science, Indiana University Purdue University, USA. Retrieved March 22, 2006, from http://www.cs.iupui.edu/uniFrame/

Czerwinski, S., Zhao, B., Hodes, T., Joseph, A., & Katz, R. (1999). An architecture for a secure service discovery service. *Proceedings of ACM Mobicom'99* (pp. 24-35). Retrieved March 22, 2006, from http://bnrg.cs.berkeley.edu/~czerwin/publications/sds-mobicom.pdf

Gandhamaneni, J. (2004). *UniFrame mobile, agent-based resource discovery system (MURDS).* Unpublished MS project, Department of Computer and Information Science, Indiana University Purdue University, USA. Retrieved March 22, 2006, from http://www.cs.iupui.edu/uniFrame/

Globus Alliance. (n.d.). Towards open grid services architecture. *Proceedings of the Open Grid Forum,* Chicago, IL. Retrieved March 22, 2006, from http://www.globus.org/ogsa/

Guttman, E. (1999). Service location protocol: Automatic discovery of IP network services. *IEEE Internet Computing, 3*(4), 71-80.

Kandagatla, C. (2003). *Survey and taxonomy of Grid resource management systems.* Retrieved March 22, 2006, from http://www.cs.utexas.edu/users/browne/cs395f2003/projects/KandagatlaReport.pdf

Magedanz, T., Bäumer, M., & Choy, S. (n.d.). *Grasshopper—A universal agent platform based on OMG MASIF and FIPA standard*s. Retrieved March 22, 2006, from http://www.cordis.lu/infowin/acts/analysys/products/thematic/agents/ch4/ch4.htm

OASIS Consortium. (2000). *UDDI technical white paper.* Retrieved March 22, 2006, from http://www.uddi.org/pubs/Iru_UDDI_Technical_White_Paper.pdf

OMG (Object Management Group). (2000). *Trading object service specification.* Retrieved March 22, 2006, from http://www.omg.org/docs/formal/00-06-27.pdf

Olson, A., Raje, R., Bryant, B., Burt, C., & Auguston, M. (2005). UniFrame: A unified framework for developing service-oriented, component-based, distributed software systems. In Z. Stojanovic & A. Dahanayake (Eds.), *Service oriented software system engineering: Challenges and practices* (pp. 68-87). Hershey, PA: Idea Group Publishing.

Raje, R. (2000). UMM: Unified Meta-object Model for open distributed systems. *Proceedings of the 4th IEEE International Conference on Algorithms and Architecture for Parallel Processing* (pp. 454-465). Los Alamitos, CA: IEEE Press. Retrieved March 22, 2006, from http://www.cs.iupui.edu/uniFrame

Raje, R., Auguston, M., Bryant, B., Olson, A., & Burt, C. (2001). A unified approach for the integration of distributed heterogeneous software components. *Proceedings of the Workshop on Engineering Automation for Software Intensive System Integration* (pp. 109-119). Monterey, CA: U.S. Naval Postgraduate School. Retrieved March 22, 2006, from http://www.cs.iupui.edu/uniFrame

Seacord, R., Hissam, S., & Wallnau, K. (1998). *Agora: A search engine for software components.* Technical Report, CMU/SEI-98-TR-011, ESC-TR-98-011, Carnegie Mellon University, USA.

Siram, N. (2002). *An architecture for discovery of heterogeneous software components.* Unpublished MS thesis, Department of Computer and Information Science, Indiana University Purdue University Indianapolis, USA. Retrieved March 22, 2006, from http://www.cs.iupui.edu/uniFrame/

Sun Microsystems. (1994). *Remote method invocation.* Retrieved March 22, 2006, from http://java.sun.com/products/jdk/rmi

Sun Microsystems. (2001). *Designing enterprise applications with the J2EE^{TM} platform.* Retrieved March 22, 2006, from http://java.sun.com/blueprints/guidelines/designing_enterprise_applications/

Waldo, J. (1999). The Jini architecture for network-centric computing. *Communications of ACM, 42*(7), 76-82.

W3C. (2004). *SOAP versions & reports.* Retrieved March 22, 2006, from http://www.w3.org/TR/soap

KEY TERMS

Active Registry: An enhanced version, capable of listening to broadcasts, of the basic registration mechanism present in a component model.

Client Query Result Retrieval Time (CQRRT): The time taken from the instant a client issues a query to MURDS until the instant the results return back to the client.

Distributed Computing System (DCS): A system made up of networked processors, each with its own memory, that communicate with each other by sending messages.

Headhunter: A critical piece of MURDS whose task is to accept queries and send mobile agents to discover the requested components, which are deployed over the network.

Internet Component Broker (ICB): A collection of services that provides a secure infrastructure for accepting incoming queries from distributed system developers, propagating the queries to the headhunters, and collecting results from the headhunters.

Mobile Agent: A software agent that can migrate from one point of connection to another on a network.

Mobile Agent-based Resource Discovery System (MURDS): An enhancement of URDS. It uses mobile agents for locating deployed services in a network and for propagating the discovery queries.

UniFrame: A unifying framework that supports a seamless integration of distributed and heterogeneous components.

UniFrame Resource Discovery System (URDS): Provides an infrastructure for proactively discovering components deployed over a network.

Mobile Business Applications

Cheon-Pyo Lee
Carson-Newman College, USA

INTRODUCTION

As an increasing number of organizations and individuals are dependent on mobile technologies to perform their tasks, various mobile applications have been rapidly introduced and used in a number of areas such as communications, financial management, information retrieval, and entertainment. Mobile applications were initially very basic and simple, but the introduction of higher bandwidth capability and the rapid diffusion of Internet-compatible phones, along with the innovations in the mobile technologies, allow for richer and more efficient applications.

Over the years, mobile applications have primarily been developed in consumer-oriented areas where products such as e-mail, games, and music have led the market (Gebauer & Shaw, 2004). According to the ARC group, mobile entertainment service will generate $27 billion globally by 2008 with 2.5 billion users (Smith, 2004). Even though mobile business (m-business) applications have been slow to catch on mobile applications for consumers and are still waiting for larger-scale usage, m-business application areas have received enormous attention and have rapidly grown. As entertainment has been a significant driver of consumer-oriented mobile applications, applications such as delivery, construction, maintenance, and sales of mobile business have been drivers of m-business applications (Funk, 2003).

By fall of 2003, Microsoft mobile solutions partners had registered more than 11,000 applications including e-mail, calendars and contacts, sales force automation, customer relationship management, and filed force automation (Smith, 2004). However, in spite of their huge potential and benefits, the adoption of m-business applications appears much slower than anticipated due to numerous technical and managerial problems.

BACKGROUND

M-business applications can be classified into two distinct categories in terms of target groups: vertical and horizontal target group (Paavilainen, 2002). Vertical targets are typically narrow user segments, such as filed service engineers or sales representatives. On the other hand, horizontal targets are a massive number of users. For example, mobile e-mail, mobile bulletin board, and mobile calendar are applications for a horizontal target group, while mobile recruitment tools, mobile sales reporting, and mobile remote control represent vertical applications (see Table 1). Generally, the goal of horizontal applications is to improve communication and streamlined processes in horizontal procedures, such as travel management and time entry. In contrast, the goal of vertical applications is to improve and solve business processes in more detailed and specific areas such as the needs of sales departments. Various vertical and horizontal applications are currently used in a number of industries. Table 2 provides examples of m-business applications in various industries.

THE IMPACTS OF MOBILE BUSINESS APPLICATIONS ON BUSINESSES

The advantages of using m-business applications are mobility, flexibility, and dissemination of m-business applications (Nah, Siau, & Sheng, 2005). Mobility allows users to conduct business anytime and anywhere, and flexibility allows users to capture data at the source or point of origin. In addition, m-business applications offer an efficient means of disseminating real-time information to a larger user population, which consequently enhances and improves customer service. According to Gebauer and Shaw (2004), users valued two

Table 1. Examples of vertical and horizontal mobile business applications (Paavilainen, 2002)

Vertical Mobile Applications	Horizontal Mobile Applications
• Mobile e-mail • Mobile bulletin board • Mobile time entry • Mobile calendar • Mobile travel management • Mobile pay slips	• Mobile recruitment tools • Mobile tools for filed engineers • Mobile sales reporting • Mobile supply chain tools • Mobile fleet control • Mobile remote control • Mobile job dispatch

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 2. Examples of various mobile business applications (Sources: Chen & Nath, 2004; Collett, 2003; Dekleva, 2004)

Hotel	• Embassy Suite: Maintenance and housekeeping crews are equipped with mobile text messaging devices, so the front desk can inform the crew of the location and nature of the repair without physically locating them. • Las Vegas Four Seasons: Customer food orders are wirelessly transmitted from the poolside to the kitchen. • Carlson hotels: Managers use Pocket PCs to access all of the information they need to manage the properties in real-time.
Hospital & Healthcare	• Johns Hopkins Hospital: Pharmacists use a wireless system for accessing critical information on clinical interventions, medication errors, adverse drug reactions, and prescription cost comparisons. • St. Vincent's Hospital: Physicians can retrieve a patient's medical history from the hospital clinical database to their PDA. • ePocrates: Healthcare professionals receive drug, herbal, and infections disease information via handheld devices.
Insurance	• Producer Lloyds Insurance: Field agents can assess the company's Policy Administration & Services System (PASS) and Online Policy Updated System (OPUS).
Government	• Public safety agencies can access federal and state database and file reports.
Manufacture	• General Motors: Workers can receive work instructions wirelessly • Celanese Chemicals Ltd.: Maintenance workers are able to arrange for repair parts and equipment to be brought to the site using wireless Pocket PCs. • Roebuck: Technicians can communicate and order parts directly from their job location instead of first walking back to their truck.
Delivery Service	• UPS & FedEx : Drivers can access GPS and other important information in real-time

things most in m-business applications use: notification, especially in connection with high mobility, and support for simple activities like tracking. The study suggested that the combination of mobility and the frequency with which each task occurred is a primary indicator of the usage of m-business applications.

M-business applications have shown significant impacts and created enormous business values. For example, m-business applications have improved operational efficiency as well as flexibility and the ability to handle situations to current operations (Chen & Nath, 2004; Gebauer & Shaw, 2004). In addition, m-business applications allow users to have access to critical information from anywhere at anytime, resulting in greater abilities to seize business opportunities.

It is very difficult to measure the direct impact of mobile business applications in *productivity* statistics, but according to an OMNI (2005) consulting report, financial services agents executed approximately 11.4% more trade options on an annualized basis with mobile business applications and achieved an average nominal improvement of 3.1% in overall portfolio performance. Also, health care and pharmaceutical filed sales representatives conducted an additional 8.3 physical briefings per week due to mobile business applications. Finally, insurance-filed claims adjusters handled an additional 7.4 claims per worker per week and improved payout ratios by an annual yield of 6.4% per adjuster using mobile business applications. Table 3 provides a list of values created by mobile business applications.

FACILITATORS AND INHIBITORS OF MOBILE BUSINESS APPLICATIONS GROWTH

Several factors are expected to contribute to the continued growth of m-business applications. Across the globe, mobile devices such as Internet-enabled mobile phones and personal digital assistants (PDAs) are gaining rapid popularity among businesses and consumers. This rapid penetration of mobile devices can provide strong support for mobile business applications. Employees' demand to access critical business processes and services from anywhere at any time is also a significant driving factor for m-business applications (Chen

Table 3. Values of mobile business applications (Sources: Chen & Nath, 2004)

Value	
Efficiency	Reduce business process cycle time
	Capture information electronically
	Enhance connectivity and communication
	Track and surveillance
Effectiveness	Reduce information float
	Access critical information anytime-anywhere
	Increased collaboration
	Alert and m-marketing campaigns
Innovation	Enhance service quality
	React to problems and opportunities anytime-anywhere
	Increase information transparency to improve supply chain
	Localize

& Nath, 2004). The traditional methods of wired communication, which have a limited reach and range, are no longer suitable for the fast-paced business environment. Finally, corporate and individual customers, who are demanding more channels for interaction and services, also contribute to the growth of m-business applications.

However, in spite of their huge potential and benefits, the adoption of m-business applications appears much slower than anticipated. Various factors have been offered as explanations for this slow growth, including the immaturity of the wireless technology, the existence of a chaotic array of competing technologies and standards, and the lack of killer applications (Chen & Nath, 2004). According to Gebauer and Shaw (2004), poor technology characteristics have inhibited application usage to a great extent. In addition, according to Nah et al. (2005), security, cost, and employee acceptance are also significant barriers of the growth of m-business applications. Companies have been concerned about the loss or theft of mobile devices, which are easily misplaced or stolen, and their likelihood to contain sensitive or confidential data that can be accessed by unauthorized persons. Huge cost is also a concern to companies. To implement mobile applications, the company must invest in mobile devices, pay service fees for wireless access, and train employees. According to Lucas (2002), some U.S. firms are spending between $5 million and $50 million for mobile business applications. Finally, employee acceptance is also a big barrier. Not every employee is willing to embrace new technology, and some employees accustomed to standard operation procedures resist adoption of m-business applications.

FUTURE TRENDS

In the future, more customized and personalized business applications will be introduced. These applications are called context-aware or situation-dependent m-business applications (Figge, 2004; Heer, Peddemors, & Lankhorst, 2003). Currently, the majority of context-aware computing has been restricted to location-aware computing for mobile applications. However, more contextual information including spatial (e.g., speed and acceleration), temporal (e.g., time of the day), environmental (e.g., temperature), and social situation (e.g., office nearby) information will be added to increase the value of mobile business applications. In context mobile business applications, the most necessary information for the user to perform tasks will be provided in advance without the user's involvement. Therefore, in most cases, the user simply presses a single button rather than making several text inputs.

However, for m-business applications to grow, current limitations in technical and managerial issues should be resolved. Current technical limitations are mainly related to mobile devices such as small multi-function keypads, less computation power, and limited memory and disk capacity (Siau, Lim, & Shen, 2001). Other technical issues such as the lack of network standards and security problems also must be resolved (Chen & Nath, 2004). In addition, a clear understanding of the value of m-business applications is also very important to grow m-business applications. The m-business development and adoption decision should always be based on clearly identified needs and business requirements (Paavilainen, 2002).

CONCLUSION

M-business applications have shown significant impacts on business processes. M-business applications not only increase productivity, but also develop new business processes that yield increased customer and job satisfaction as well as competitive advantage. In the future, richer and more efficient m-business applications will be introduced to attract more businesses. However, current technical and managerial limitations should be resolved to support continued growth of m-business applications. Especially, it is very important to understand the fundamental value derived from m-business applications before developing and adopting them.

REFERENCES

Chen, L.-D., & Nath, R. (2004). A framework for mobile business applications. *International Journal of Mobile Communications, 2,* 368-381.

Collett, S. (2003). Wireless gets down to business. *Computerworld, 37*(18), 31.

Dekleva, S. (2004). M-business: Economy driver or a mess? *Communications of the Association for Information Systems, 13,* 111-135.

Figge, S. (2004). Situation-dependent services: A challenges for mobile network operators. *Journal of Business Research, 57*(12), 1416-1422.

Funk, J. (2003). *Mobile disruption: Key technologies and applications that are driving the mobile Internet.* New York: John Wiley & Sons.

Gebauer, J., & Shaw, M.J. (2004). Success factors and impacts of mobile business applications: Results from a mobile e-procurement study. *International Journal of Electronic Commerce, 8*(3), 19-41.

Heer, J.D., Peddemors, A.J.H., & Lankhorst, M.M. (2003). *Context-aware mobile business applications.* Retrieved October 29, 2005, from https://doc.telin.nl/dscgi/ds.py/Get/File-25810/coconet.pdf

Lucas, M. (2002). Wireless financial apps grow slowly. *Computerworld, 36,* 14.

Nah, F.F.-H., Siau, K., & Sheng, H. (2005). The value of mobile applications: A utility company study. *Communications of the ACM, 48,* 85-90.

Omni. (2005). *Study finds 13.4 percent increase in worker productivity.* Retrieved October 10, 2005, from http://newsroom.cisco.com/dlls/2005/prod_020905.html

Paavilainen, J. (2002). *Mobile business strategies.* London: Wireless Press.

Siau, K., Lim, E.P., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3), 4-13.

Smith, B. (2004). Business apps: Going for the tried and true. *Wireless Week, 10,* 22.

KEY TERMS

Horizontal Mobile Business Application: Mobile business application developed for a massive number of users to improve communication and streamline processes.

Location-Aware Computing: The capability of computing to recognize and react to location context. Global Positioning System (GPS) is the most widely known location-aware computing system.

Mobile Business Application: Mobile application used to perform business tasks such as sales force automation, customer relationship management, and filed force automation.

Situation-Dependent Mobile Application: Mobile application using various contextual information such as spatial, temporal, environmental, and social.

Vertical Mobile Business Application: Mobile business application developed for a specific target group such as filed service engineers and sales representatives.

Mobile Cellular Traffic with the Effect of Outage Channels

Hussein M. Aziz Basi
Multimedia University, Malaysia

M. B. Ramamurthy
Multimedia University, Malaysia

INTRODUCATION

The designer of the cellular network must evaluate the possible configurations of the system components and their characteristics in order to develop a system with greater efficiency. This article studies the grade of service (GOS) degradation in the presence of outage for a mobile cellular network where the number of channels in outage can be used as an indicator of the traffic load for two models, namely fixed outage rate and traffic dependent outage rate. The performance parameters considered for this article are: the probability of delay, waiting time for priority and non-priority calls, mean waiting time, and priority gain; each is estimated for both models. The system is evaluated and compared under different conditions.

BACKGROUND

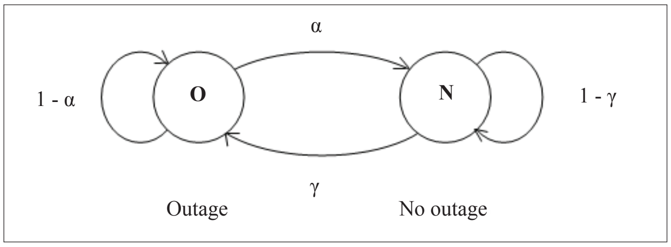
The mobile user behavior has a higher traffic impact (in both space and time) than in the fixed network line. The call initiation sites are scattered and dynamically changing over a geographical area, while the bandwidth associated with a connection may have to be provided to different sites throughout the call; the radio signal will change from one cell to another following the user call movement, and in such environments, the efficient allocation of wireless channels for communication sessions is of vital importance as the bandwidth allotted for cellular communication is limited. The number of wireless communication service users as well as the frequency of the available services increased with an unexpected rate. The analysis of traffic deployed with wireless communication networks is important for determining the operation for a mobile user's status. Mobile cellular traffic varies greatly from one period to another and not in any uniform manner, but according to the cellular user's needs. Teletraffic theory is used to specify the methods to ensure that the actual GOS is fulfilling the requirements. The calls in the cellular network are made by individual customers according to their habits, needs, and so forth, and the overall pattern of calls will vary throughout the day. The cellular network equipment should be sufficient in quantity to cope satisfactorily for the period of maximum demand in the busy hour, depending on availability of free channel. In order to determine the optimal channel loading, it is necessary to relate the GOS to traffic characteristics. Traffic modeling is necessary for cellular network provisioning, for predicting utilization of cellular network resources, and for cellular network planning and developments (Vujicic, Cackov, Vujicic, & Trajkovic, 2005) to specify emergency actions when systems are overloaded or technical faults occur. GOS could be defined as the number of unsuccessful calls relative to the total number of attempted calls (Nathan, Ran, & Freedman, 2002; Zhao, Shen, & Mar, 2002; Hong, Malhamd, & Gerald, 1991).

Voice traffic network has been modeled by the Erlang C formula (Yacoub, 1993), which is used in cases where all users have access to all channels in the mobile network and where there are a large number of users using the available channels (Nathan et al., 2002). The number of required channels is used as a fraction of user traffic intensity and desired GOS. The GOS in a cellular system is affected not only by the system's traffic but also by co-channel interference. The cellular system presence of co-channel interference can cause the carrier-to-interference ratio (C/I) to drop below a specified threshold level (Annamalai, Tellambura, & Bhargava, 2001; Aguirre, Munoz, Molina, & Basu, 1998; Yang & Alouini, 2002, 2006; Zhang, 1996**); s**uch an event is known as outage. In some cases, outage can cause the loss of the communication system.

Aguirre et al. (1998) estimated the effect of an outage channel for many models where there are no available channels and the call is blocked or dropped. In this case they did not consider the aspect of buffering the dropped calls (outage calls) and the mean waiting time with priority calls. Another researcher evaluated the performance of mobile systems with priority concept, where no channel is available when the call is queued through to when the available channel has been assigned, and the priority calls are placed in a queue before all non-priority calls but never interrupt a call in progress (Barcelo & Paradells, 2000); however they did not consider the concept of outage.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Outage parameter



The major contributions of this article are to analyze the performance of a mobile communication system including GOS degradation due to the outage, when the calls as well as the outage channels in the cellular system are queued in the same buffer according to their priority, and for two different models to evaluate the performance of outage channel on the cellular network.

THE OUTAGE PARAMETERS

When the C/I is dropped below a certain quality threshold (ϑ) in a given channel, it becomes unusable and it affects the GOS in the cell. While two subscribers are communicating in the cellular network, the user could experience an absence of the desired signal and some noise or crosstalk. Even if link outages are very short, they collectively degrade the system performance, although they may not be individually recognized. Generally only outages listing longer than tens of milliseconds are recognized and can cause the dropout of the communication (Caini, Immovilli, & Merani, 2002). When the new calls come to the cellular system (by arrival rateλ) and there is a free channel in the cell, the call will engage one of the free channels and the channel becomes busy. The channel can go into outage with outage arrival rate γ, and the outage channel may recover by the outage recovery rate α. Thus, the numbers of available channels for service become a random variable due to the stochastic nature of the outage.

A channel from a normal working condition may become unavailable (or move into the outage state) due to drop in C/I. Thus, the two-state simple model shown in Figure 1 can represent its behaviors. The state $\boldsymbol{O}$ represents a channel in outage, while state $\boldsymbol{N}$ represents a state in it normal condition. The parameters γ and α represent failure and recovery rates. These rates can be represented in terms of steady-state probabilities $\boldsymbol{O}$ and $\boldsymbol{N}$ by the following analysis.

The probabilities of being in states $\boldsymbol{O}$ and $\boldsymbol{N}$ are:

$$\boldsymbol{O} = \gamma\boldsymbol{N} + (1-\alpha)\boldsymbol{O} \tag{1}$$

$$\boldsymbol{N} = (1-\gamma)\boldsymbol{N} + \alpha\boldsymbol{O} \tag{2}$$

In addition, they satisfy $\boldsymbol{O}+\boldsymbol{N} = 1$. After solving this system of equations, the following is obtained:

$$\boldsymbol{O} = \frac{\gamma}{\gamma+\alpha} \tag{3}$$

$$\boldsymbol{N} = \frac{\alpha}{\gamma+\alpha} \tag{4}$$

By sorting out $\gamma + \alpha$ in both previous equations and equalizing them, it is found that

$$\alpha = \boldsymbol{N}\gamma / \boldsymbol{O} \tag{5}$$

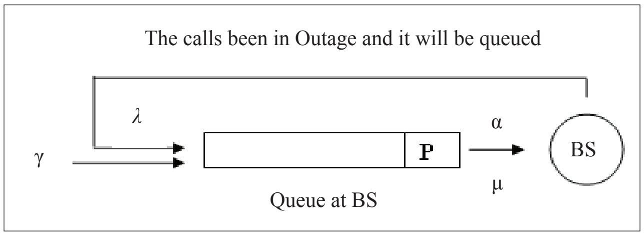
With the above equation, the outage arrival rate γ or the outage recovery rate α, assuming one of them, a value of the outage probability can be obtained. The relations for the outage γ and α are used to find the probability of delay for a different cellular system under different conditions. The design is extended for the normal cellular system by considering the outage channels where the outage channel calls as well as the normal incoming calls are queued in the same buffer as shown in Figure 2.

Queues occur wherever an unbalance occurs between requests for a limited resource and the ability of a service facility to provide that resource. The size of the buffer depends on the amount of the resource available and the demand for it by subscriber. The most common service discipline in real life is called first in first out (FIFO) or first come first served (FCFS); non-preemptive priority calls are used in this article where the priority calls have been affected by outage and thus the priority calls will be re-queued in the head of the buffer as it is assigned as high priority. Queuing of new calls and waiting for requests to be served can generally improve channel utilization at the expense of time spent in the queue.

PRESENT MODELS

Outage can cause the loss of the communication system and affects the GOS. Study of the GOS degradation due to

Figure 2. Queuing the outage and the normal calls



BS: Base Station, P: Priority calls, λ: Arrival rate, μ: Departure rate, γ : Outage arrival rate, α : Outage recovery rate

outage will be useful in evaluating the system performance. For a mobile communication system with channels in outage, a modification for the Erlang C formula is proposed by the authors for two cases, namely fixed outage rate and traffic dependent outage rate model (Basi & Murthy, 2004, 2005). In fixed outage rate model, the outage arrival rate (γ) is independent of the state, while the outage recovery rate (α) increases with the number of channels in outage. In a traffic-dependent outage rate model, both outage arrival rate (γ) and outage recovery rate (α) are state dependent. The most important queuing system, called Erlang C, has been widely used to evaluate the queuing system behavior as shown below (Yacoub, 1993). This equation is known as the Erlang C formula:

$$C(N,A) = \frac{A^N}{N!}\frac{1}{1-A/N}p_0 \tag{6}$$

where

$$P_0 = \left[\sum_{k=0}^{N-1}\frac{A^k}{k!} + \frac{A^N}{N!}\frac{1}{1-A/N}\right]^{-1} \tag{7}$$

The modified formula is carried out in this article to evaluate the probability of delay for both models. Then using this probability of delay, the mean waiting time is calculated. Considering a priority option for some calls, the priority gain is estimated in view of the modification to the Erlang C formula. Thus analysis is carried out for two situations, namely priority and non-priority cases. The modification formula for the Erlang C model is proposed by Basi and Murthy (2004, 2005)**.** The effective traffic intensity is taken to be $A_e = A + A_o$, replacing A in the Erlang C formula. In a system when k channels are in outage, the available channels will be N-k. Thus using the effective rates and number of channels as (N-k) instead of N in Erlang's C formula, the modified expression for probability of delay is:

$$P_D = C(N-k, A_e) = \frac{A_e^{\,N-k}}{(N-k)}\frac{p_0}{1-\dfrac{A_e}{(N-k)}} \tag{8a}$$

where

$$p_o = \left[\sum_{i=0}^{N-k-1}\frac{A_e^{\,i}}{i!} + \frac{A_e^{\,N-k}}{(N-k)}\frac{1}{1-\dfrac{A_e}{N-k}}\right]^{-1}$$

where N > k and $A_o < 1$ (8b)

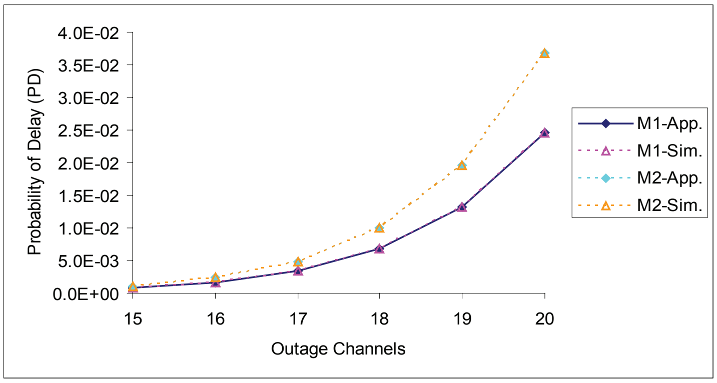
These expressions are used in the present work. The probability of delay is estimated with varying call duration, number of calls/hour, and number of outage channels. The probability of delay is calculated according to the proposed formula, and it is used to evaluate the waiting times when a channel is in outage. The total duration (d) of call is known as call duration plus the recovery rate time; using this concept to calculate mean waiting time of priority and non-priority calls, the relations of Barceló and Paradells (2000) are used with modification, where N is replaced by N-k and d is taken as $d_r + \alpha$; d_r as duration time and α as outage recovery rate used in the modified relations will be:

$$WT_1 = \frac{(P_D * d)}{(N-k)(1-p\rho)} \tag{9}$$

$$WT_2 = \frac{(P_D * d)}{(N-k)(1-p\rho)(1-\rho)} = \frac{WT_1}{(1-\rho)} \tag{10}$$

WT_1 and WT_2 are the waiting time for calls with priority and non-priority (regular calls) respectively, ρ is overall system load, and p is the priority proportion. The load due to priority calls will be ρ x p. The evaluation conditions of channel system are heavy traffic (high ρ) and low priority propagation (low p) to maintain the effectiveness of the

Figure 3. P_D vs. outage channels (k) where $\alpha = 0.2$ and $\gamma = 0.00664$



priority system as statues (Barceló & Paradells, 2000). It can easily be checked that the following relation holds for the average mean waiting time for all calls:

$$W_m = p\, WT_1 + (1-p)\, WT_2 \tag{11}$$

The priority gain is convenient ration. It is the quotient between the mean waiting time for all calls (as if there was no priority) and mean waiting time for priority calls:

$$P_G = W_m / WT_1 \tag{12}$$

The parameters are calculated using the original Erlang C formula, and unmodified relations of Barceló and Paradells (2002) are given as results for normal system. The simulation system has been designed to evaluate the above models, where the coming calls as well as the calls of outage channels in the cellular system are queued with consideration of the following assumption (Hussein & Murthy, 2006):

1. All the channels are fully available for servicing calls until all channels are occupied.
2. The offered traffic is uniformly distributed in the cell.
3. The number of subscribers is assumed infinite.
4. The call is initiation as a Poisson process with a mean call arrival of λ calls/hour.
5. The call holding time is exponentially distributed with a mean of 120 s.
6. The threshold level for new calls >19dB and for dropped calls <17.3 dB.
7. The interference channel (outage channels) should be limited.
8. The outage recover rate is 0.00664.
9. The buffer is assumed infinite.
10. Sorting the calls according to their priority without interrupting calls in progress.

The analytical as well as the simulation results are obtained for these parameters and plotted as graphs for the two proposed models.

RESULTS AND DISCUSSION

The results are presented in tables and plotted as graphs for the two models. The testing is carried out with 40 channels and a call rate of 360 calls per hour, while the outage arrival rate is 0.00664. From the results of both models reported, it is found that probability of delay (P_D) is much lower in the first model (fixed outage model) than the second model (traffic-dependent outage rate model). This means that the call has to wait for shorter time in the buffer in the case of a fixed outage rate model, as shown in Figure 3 and Table 1.

The waiting time for priority calls (WT_1) and non-priority calls (WT_2) for the first model is less than those of second model under different conditions for different priority call percentages as shown in Figures 4 and 5. The mean waiting time (W_m) for model one is less than in model two as shown in Figure 6; this means that the time for the call to wait in the first model is less than that in the second model. Figure 7 shows that the priority gain (P_G) is higher in case of the second model for heavy load and low priority call percentage.

The comparison of the two proposed models with the Erlang C model and the normal system (Barceló & Paradells, 2002) for the waiting time and the priority gain are present with the value of $k = 1$, as given in Tables 2, 3, 4, and 5. The P_D for the proposed models gives a higher delay than Erlang C because of the effect of the channel in outage on the proposed models as shown in Table 2. With the WT_1, WT_2, and W_m, it is found that the waiting time for the proposed models is higher than the normal system as given in Tables 3 and 4. For P_G, it is found that proposed models are given a higher delay for heavy load and a lower delay for priority

Table 1. Comparison of P_D *between model one and two,* $\alpha = 0.2$ & $\gamma = 0.00664$

k	Model-1(A)	Model-1(S)	Model-2(A)	Model-2(S)
0	1.58116E-10	1.58115E-10	1.58116E-10	1.58115E-10
1	5.34530E-10	5.34527E-10	5.36156E-10	5.36152E-10
2	1.76274E-09	1.76273E-09	1.78350E-09	1.78349E-09
3	5.66726E-09	5.66722E-09	5.81274E-09	5.81271E-09
4	1.77525E-08	1.77524E-08	1.85380E-08	1.85379E-08
5	5.41465E-08	5.41462E-08	5.77766E-08	5.77763E-08
6	1.60702E-07	1.60701E-07	1.75742E-07	1.75741E-07
7	4.63780E-07	4.63777E-07	5.21010E-07	5.21007E-07
8	1.30057E-06	1.30056E-06	1.50338E-06	1.50337E-06
9	3.54127E-06	3.54125E-06	4.21635E-06	4.21633E-06
10	9.35532E-06	9.35526E-06	1.14773E-05	1.14772E-05
11	2.39600E-05	2.39598E-05	3.02800E-05	3.02799E-05
12	5.94415E-05	5.94412E-05	7.73161E-05	7.73157E-05
13	1.42728E-04	1.42727E-04	1.90794E-04	1.90793E-04
14	3.31424E-04	3.31422E-04	4.54396E-04	4.54394E-04
15	7.43628E-04	7.43624E-04	1.04302E-03	1.04302E-03
16	1.61094E-03	1.61093E-03	2.30458E-03	2.30457E-03
17	3.36691E-03	3.36689E-03	4.89585E-03	4.89583E-03
18	6.78465E-03	6.78461E-03	9.99031E-03	9.99027E-03
19	1.31747E-02	1.31746E-02	1.95669E-02	1.95668E-02
20	2.46444E-02	2.46443E-02	3.67676E-02	3.67675E-02

Figure 4. WT_1 *vs. the percentage of priority calls (p) where* $k = 20$, $\alpha = 0.2$, *and* $\gamma = 0.00664$

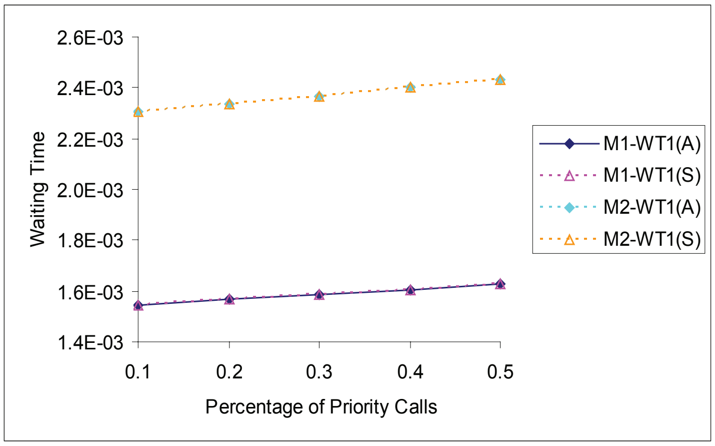


Figure 5. WT_2 vs. the priority call percentage where $k = 20$, $\alpha = 0.2$, and $\gamma = 0.00664$

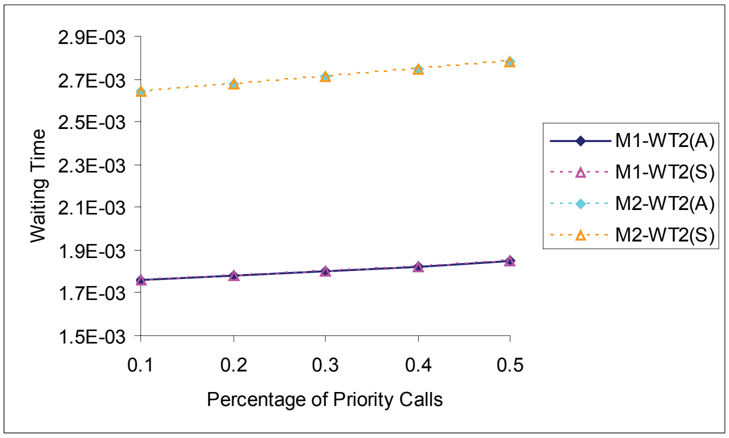


Figure 6. W_m vs. k with $\alpha = 0.2$ and $\gamma = 0.00664$

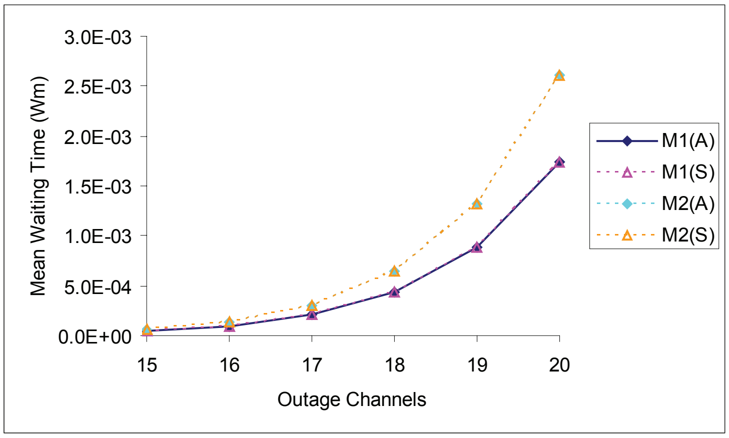


Figure 7. P_G vs. priority percentage with $k = 20$, $\alpha = 0.2$, and $\gamma = 0.00664$

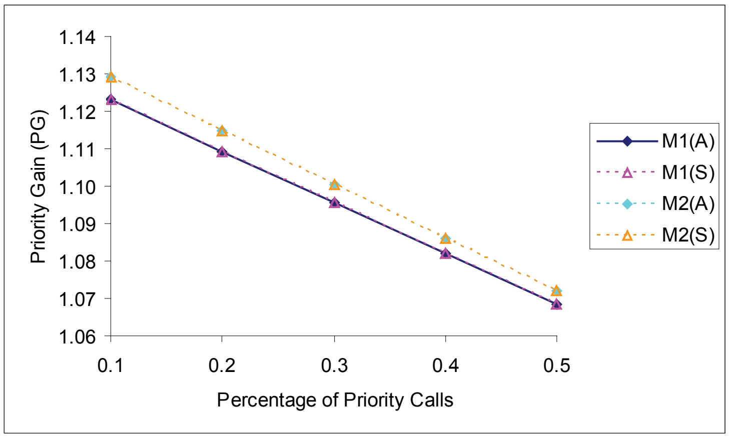


Table 2. Comparison of P_D between model one, two, and the Erlang C model; $\alpha = 0.2$, $\gamma = 0.00664$, and k=1

Channels	Erlang C	Model 1	Model 2
35	5.345E-08	1.587E-07	1.591E-07
36	1.756E-08	5.359E-08	5.373E-08
37	5.620E-09	1.761E-08	1.766E-08
38	1.752E-09	5.635E-09	5.651E-09
39	5.329E-10	1.758E-09	1.763E-09
40	1.581E-10	5.345E-10	5.362E-10
41	4.580E-11	1.586E-10	1.591E-10
42	1.296E-11	4.595E-11	4.610E-11
43	3.583E-12	1.300E-11	1.305E-11
44	9.686E-13	3.595E-12	3.608E-12
45	2.562E-13	9.721E-13	9.756E-13

Table 3. Comparison of waiting times between model one, two, and normal system; $\alpha = 0.2$, $\gamma = 0.00664$, and k = 1

p	WT_1	M1- WT_1	M2- WT_1	WT_2	M1- WT_2	M2- WT_2
0.1	4.801E-12	1.667E-11	1.673E-11	5.456E-12	1.895E-11	1.901E-11
0.2	4.860E-12	1.688E-11	1.693E-11	5.523E-12	1.918E-11	1.924E-11
0.3	4.921E-12	1.709E-11	1.714E-11	5.592E-12	1.942E-11	1.948E-11
0.4	4.983E-12	1.731E-11	1.736E-11	5.662E-12	1.967E-11	1.973E-11
0.5	5.046E-12	1.753E-11	1.758E-11	5.734E-12	1.992E-11	1.998E-11

Table 4. Comparison of mean waiting time between model one, two, and normal system; $\alpha = 0.2$, $\gamma = 0.00664$, k =1, and priority percentage is 0.3

Channels	W_m	Model 1- W_m	Model 2- W_m
35	2.082E-09	6.376E-09	6.392E-09
36	6.652E-10	2.091E-09	2.097E-09
37	2.071E-10	6.681E-10	6.699E-10
38	6.289E-11	2.080E-10	2.086E-10
39	1.863E-11	6.318E-11	6.336E-11
40	5.390E-12	1.872E-11	1.878E-11
41	1.523E-12	5.416E-12	5.434E-12
42	4.207E-13	1.531E-12	1.536E-12
43	1.136E-13	4.228E-13	4.243E-13
44	3.002E-14	1.142E-13	1.146E-13
45	7.762E-15	3.018E-14	3.029E-14

Table 5. Comparison of priority gain between model one, two, and normal system; $\alpha = 0.2$, $\gamma = 0.00664$, and $k = 1$

p	P_G	Model 1- P_G	Model 2 - P_G
0.1	1.12273	1.12274	1.12276
0.2	1.10909	1.10910	1.10912
0.3	1.09545	1.09547	1.09548
0.4	1.08182	1.08183	1.08184
0.5	1.06818	1.06819	1.06820

call percentage as in Table 5. The results of the mathematical model are indicated by (A) and those of simulation by (S). M1 refers to model one (fixed outage model) and M2 to model two (traffic-dependent outage rate model).

CONCLUSION

The mobile cellular system has appeared more recently as a consequence of the high demand for mobile services; the Erlang C model is used in cases where all users have access to all channels in the mobile network and where there are a large number of users using the available channels. Co-channel interference can cause the carrier-to-interference ratio (C/I) to drop below a specified threshold level, and such an event is known as outage. The calls of outage channels are queued along with normal arriving calls into the same buffer. The call duration, number of calls per hour, and number of channels in outage affect the probability of delay as observed. The study helps in understanding the performance of a mobile link and may help in deciding the number of channels for given traffic. With increasing probability of delay to meet given traffic demands, one has to select a memory of suitable size so that all waiting calls can be queued up. According to the scheme proposed by authors, no call is lost, but they may be delayed. The above results can be successfully used in designs of cellular systems to decide the number of channels needed for satisfactory, reliable operation of the cellular system.

REFERENCES

Annamalai, A., Tellambura, C., & Bhargava, V. K. (2001). Simple and accurate methods for outage analysis in cellular mobile radio system—a unified approach. *IEEE Transactions in Communications, 49*(2), 303-308.

Aguirre, A., Munoz, D., Molina, C., & Basu, K. (1998). Outage—GOS relationship in cellular systems. *IEEE Communications Letters, 2*(1), 5-7.

Barcelo, F., & Paradells, J. (2000, September). Performance evaluation of Public Access Mobile Radio (PAMR) systems with priority calls. *IEEE Proceedings of the 11th PIMRC* (pp. 979-983), London.

Basi, H. M. A., & Murthy, M.B.R. (2004). A simple scheme for improved performance of fixed outage rate cellular system. *American Journal of Applied Sciences, 1*(3), 190-192.

Basi, H. M. A., & Murthy, M. B. R. (2005). Improved performance of traffic dependent outage rate cellular system. *Journal of Computer Sciences, 1*(1), 72-75.

Basi, H. M. A., & Murthy, M. B. R. (2006). The simulation study on the effect of outage channels on mobile cellular network. *International Journal of Computer Science & Network Security, 6*(4), 146-150.

Caini, C., Immovilli, G., & Merani, M. L. (2002). Outage probability for cellular mobile radio systems: Simplified analytical evaluation and simulation results. *Electronics Letters, 28*(7), 669-671.

Hong, H.H., Malhamd, R., & Chen, G. (1991, May 19-22). Traffic engineering of trunked land mobile radio dispatch system. *Proceedings of the 41st IEEE Vehicular Technology Conference: Gateway to the Future Technology in Motion* (pp. 251-256).

Nathan, B., Ran, G., & Freedman, A. (2002). Unified approach of GOS optimization for fixed wireless access. *Vehicular Technology, IEEE Transactions, 51*(1), 200-208.

Vujicic, B., Cackov, N., Vujicic, S., & Trajkovic, L. (2005). Modeling and characterization of traffic in public safety wireless networks. *Proceedings of SPECTS 2005* (pp. 214-223), Philadelphia, PA.

Yang, L., & Alouini, M.-S. (2006). Performance comparison of different selection combining algorithms in presence of co-channel interference. *Vehicular Technology, IEEE Transactions, 55*(2), 559-571.

Yang, L., & Alouini, M.-S. (2002). Outage probability of dual-branch diversity system in presence of co-channel interference. *IEEE Transactions on Wireless Communication, 2*(2), 310-319.

Yacoub, M. D. (1993). *Foundations of mobile radio engineering.* CRC Press.

Zhang, Q. T. (1996). Outage probability in cellular mobile radio due to Nakagami signal and interferers with arbitrary parameters. *Vehicular Technology, IEEE Transactions, 45*(2), 364-372.

Zhao, D., Shen, X., & Mar, J. W. K. (2002). Performance analysis for cellular system supporting heterogeneous services. *Proceedings of ICC 2002* (vol. 5, pp, 3351-3355).

Mobile Commerce

JiaJia Wang
University of Bradford, UK

Pouwan Lei
University of Bradford, UK

INTRODUCTION

The rapid development and deployment in wireless networks and mobile telecommunication systems are leading to a phenomenal growth of innovative and intelligent mobile applications generally referred to as mobile commerce (m-commerce). Mobile devices like the mobile phone become a necessity for everyone. M-commerce makes networks more productive by seamlessly bringing together voice, data communication, and multimedia services. There is an increasing demand in mobile applications or m-commerce. The objective of this short article is to discuss the reasons for the growth of m-commerce. First, variety of wireless and mobile telecommunication technologies will be reviewed. Second, the evolution of m-commerce application architecture will be studied. Third, we will examine the landscape of m-commerce. Finally, we conclude the article.

BACKGROUND

The recent phenomenal convergence of the Internet and mobile telecommunication has accelerated the demand for "Internet in the pocket" on light, low-cost terminals, as well as for radio technologies that boost data throughput and reduce the cost per bit. This trend to higher data rates over wireless networks will culminate in the introduction of 3G IMT-2000 (International Mobile Telecommunications-2000) systems. This revolution continues to 3.5G, which is HSDPA (High-Speed Downlink Packet Access) spreading in Europe and Japan currently, and further will get to 3.75G-HSUPA for solving uplink problems. In addition to these wide area cellular networks, a variety of wireless transmission technologies are being deployed, including DAB (Digital Audio Broadcast), DVB (Digital Video Broadcast), and DMB (Digital Multimedia Broadband) for wide area broadcasting; LMDS (Local Multipoint Distribution System) and MMDS (microwave multipoint distribution system) for fixed wireless access; and IEEE 802.11b, a, g, h, and the new standard i for WLAN (Wireless Local Area Networking), as well as WiMAX (Worldwide Interoperability for Microwave Access) extending from the enterprise world into the public and residential domains.

M-commerce, which refers to access to the Internet via a handheld device such as a cell phone or a PDA, is becoming a leading driver for the successful rollout of the current cellular systems, and will influence the relations between existing and emerging players (Paavilainen, 2001). It is expected to be one of the most important applications for nearly all social classes, as the UMTS Forum predicted the significant potential of the mobile Internet for m-commerce in 3G with the expectation about 50% of mobile subscribers (UMTS Forum, 2003), with a further 1.5 billion mobile users worldwide. The target m-commerce applications imaginable today are ranging from telemetry and credit card applications to electronic postcards, Web browsing, audio or video on demand, and even videoconferences. This will result in an estimated m-commerce global revenue of US$88 billion, and the ticket purchased and phone-based retail POS sales will result US$39 billion and US$299 million respectively in 2009 (Juniper Research, 2004).

This rapid development of m-commerce technologies has opened up hitherto unseen business opportunities. It has increased an organization's ability to reach its customers regardless of location and distance, and has also been successful to a certain extent in creating a consumer demand for more advanced mobile devices with interactive features. While the distinctive e-commerce is characterized by e-marketplaces, an explosion in m-commerce innovative applications has presented the business world with a fresh set of strategy based on personalized and location-based services (Buvat, 2005).

THE EVOLUTION OF M-COMMERCE ARCHITECTURE

M-commerce is enabled by a combination of technologies such as networking, embedded systems, databases, and security. Mobile hardware, software, and wireless networks enable m-commerce systems to transmit data more quickly, locate a user's position more accurately, and conduct business with better security and reliability. In this section, three areas of technologies that are fundamental for m-commerce will be examined which are wireless networks, wireless protocol(s), and mobile devices.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Wireless Networks

Wireless networks provide the backbone of m-commerce activities. The evolution of wireless networks continued with the implementation of 2G (Second-Generation) systems such as TDMA (Time Division Multiple Access), CDMA (Code Division Multiple Access), and GSM (Global System Of Mobile Communication), which were also used primarily for voice applications, with the exception of the SMS (Short Message Service) capability offered by the GSM network. An upgrade of the 2G networks is referred to as 2.5G wireless networks such as high-speed circuit-switched data, GPRS (General Packet Radio Service), and EDGE (Enhanced Data Rates For Global Evolution). Being either circuit-switched or packet-switched, these networks are primarily intended to allow for increases in data transmission rates and, in the case of packet-switched networks, an "always-on" connection.

3G networks are commonly referred as IMT-2000 on a global scale. Along with voice functionality, 3G networks support higher-speed transmission for high-quality audio and video enabled through high-bandwidth data transfers, as well as provide a global "always on" roaming capability. Better modulation methods and smart antenna technology are two of the main research areas that enable fourth-generation wireless systems to outperform third-generation wireless network (PriceWaterhouseCoopers, 2001).

Wireless Protocol(s)

Wireless networks are evolving, similar to the communication protocols; WAP and iMode are the two main wireless protocols that are implemented in m-commerce. The following "information exchange technology" for these two protocols is described:

- Hyper-Text Markup Language (HTML) is not a suitable format for information exchange in the wireless domain, while the compact version of HTML, known as cHTML, has been used in the NTT DoCoMo's iMode services.
- eXtensible Markup Language (XML) is a meta-language, designed to communicate the meaning of the data through a self-describing mechanism. It tags data and puts content into context, therefore enabling content providers to encode semantics into their documents. For XML-compliant information systems, data can be exchanged directly, even between organizations with different operation systems and data models, as long as the organizations agree on the meaning of the data they exchange.
- Wireless Markup Language (WML), which has been derived from XML, has been developed especially for WAP (Wireless Application Protocol). It allows information to be represented as cards suitable for display on mobile devices. So WML is basically to WAP what HTML is to the Internet.

Of course, iMode is a serious competitor of WAP 2.0 (NTTCoCoMo, 2005). It has been suggested that WAP may push ahead of iMode in popularity because WAP has a large community of developers, whereas the tightly NTT-controlled iMode may be stifled by lack of development blood (Frank, 2001). As iMode evolves towards support of XHTML and TCP (Transmission Control Protocol), with the current WAP evolution, these two technologies will probably converge. It has been rumored that the iMode supporters are evolving their platforms to support WAP users by enabling WAP phones to access iMode content. This is being done in Japan, and it is one way for iMode manufacturers and service providers to sell more equipment and services. By enabling a WAP user to get iMode content, an iMode service provider could use the product as a way of convincing the WAP user to buy his or her primary service from the iMode carrier. More than likely, a gateway function will be used to act as a mediation and conversion access point.

CHTML will likely become the common markup language for both iMode and WAP. XHTML is a combination of HTML and XML, and the combined format will define the data and the presentation of the data. This convergence for the technologies will create more opportunities to content providers and the Internet industry between the wireless Internet and the wired Internet, which in turn can offer more applications to m-commerce users and further expand the subscriber base in order to grow the revenue stream.

SMS enables sending and receiving text messages to and from mobile phones. Up to 160 alphanumeric characters can be exchanged in each SMS message. Widely used in Europe, SMS messages are mainly voicemail notification and simple person-to-person messaging. It also provides mobile information services, such as news, stock quotes, sports, weather, SMS chat, and downloading of ringing tones.

In mobile communication, knowledge of the physical location of a user at any particular moment is central to offering relevant service. Location identification technologies are important to certain types of mobile commerce applications, particularly those whose content is varied depending on location. GPS (Global Positioning System), a useful location technology, uses a system of satellites orbiting the earth.

Mobile Terminal(s)

The development of mobile terminals is partly dependent on the evolution of the networks. Bandwidth is an advanced feature, while it is not the only feature that narrows down potential applications. Network-based location services are also dependent on the equipment installed by the mobile operator. Location technologies are especially important with the evolution of car navigation systems, which use network and satellite-dependent positioning. Mobile terminals inside the car are able to use both technologies in order to provide driving directions and information on special points of interest.

Another factor affecting the evolution of mobile handsets is consumer adoption. It remains to be seen how the advanced features are welcomed by the end users. For example, consumers in Europe are more concerned about ease-of-use of their handsets, while Asian consumers are more concerned about the appearance and the size of their mobile phones, and more and more members of the younger generation already regard their mobile phones as a fashion statement.

The evolution of mobile terminals will be characterized by customer segmentation. Handsets focusing on a narrow target group, such as teens, construction workers, or business professionals, have specific requirements in terms of functions and applications. Business professionals require efficient time management and team working capabilities. Teens may choose a handset with a built-in game console. Construction workers may rely on a water-resistant phone covered with rubber. Therefore, customer segmentation will be a crucial part of the future, and device manufactures have to develop several models in order to stay in the business (Paavilanimen, 2001).

As can be seen, consumer electronics and mobile communication come closer to each other by integrating new technologies with handsets. Mobile handset owners can already use their device as a calculator, MP3 player, radio, remote controller, game console, and digital camera. Naturally, devices with the biggest potential are those integrating a mobile phone with another mobile device such as a digital camera or game console. In this way, the portability of the two devices is used to create totally new service concepts.

Now, with the emergence of data services, it is likely that the size of a mobile terminal is going to be increased, as the use of Internet applications requires a bigger screen and more flexible character input methods. A 3G consumer in the UK said, “It was like going back 10 years to when mobiles where the size of a brick” (BBC News, 2004a). For this reason, some mobile carriers like AT&T Wireless now provide users with shortcuts that allow the consumer to access Web content and services through voice-activated dialing (BBC News, 2004b). Most researchers are researching and developing to compensate this defect. The latest news shows that V920, the “world first” video eyewear, can be used as a portable high-resolution display or as the ultimate viewer in the rapidly growing mobile video markets with portable DVD players, “in-car” video systems, video-enabled cell phones, game consoles, and the new personal digital media/video players. This revolutionary device overcomes the limitations of traditional direct view displays and creates big-screen images from micro displays, providing users with an unparalleled solution for mobile entertainment and information applications (3G Newsletter, 2005).

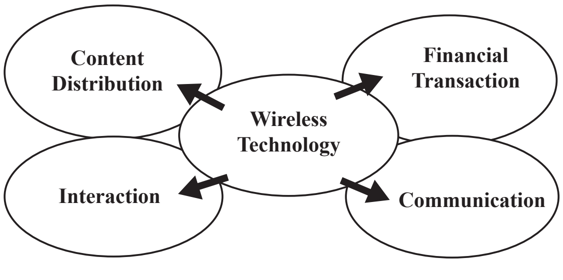
THE LANDSCAPE OF M-COMMERCE APPLICATIONS

The evolution of m-commerce applications will be driven by the user’s preference for new high-speed services and their demands, while on the move, to replicate their experience of broadband at home and work. To analyze the impact of the future m-commerce applications is challenging. All of the potential services are unforeseen, difficult to say which are really going to be the “killer applications.” For this reason the selected approach has been to broaden the granularity and use a classification that can help in assessing the penetration, usage, bandwidth, and other requirements, and thus revenue potential of the forthcoming m-commerce applications.

M-commerce operation modes can be generalized in four categories: (1) content distribution mode, (2) financial transaction mode, (3) interaction mode, and (4) communication mode—all described in detail as follows. Also, some prototypes are shown in Figure 1.

Content distribution services are concerned with real-time information notification (e.g., bank overdraft) and using positioning systems for intelligent distribution of personalized information by location (e.g., selective advertising of locally available services and entertainment). Real-time information such as news, traffic reports, stock prices, and weather forecasts can be distributed to mobile phones via the Internet. The information is personalized to users’ interests. Users also can retrieve local information such as restaurant and shopping information, as well as traffic reports. Content distribution services with a greater degree of personalization and localization can be effectively provided through a mobile portal (Tsalgatidou & Veijalainen, 2000). Localization means to supply information relevant to the current location of the user. A user’s profile—such as past behavior, situation, and location—should be taken into account for personalization and localized service provision. Notification can be sent to the mobile device too.

Figure 1. Mobile commerce operation modes



In the financial transaction mode, companies use the wireless Internet to run business transactions. M-commerce consumers can browse through the catalog and order products online. Although there are still some hidden obstacles such as transaction security, speed, and ease of use, it seems that most companies are likely to benefit directly from transactions on the wireless Internet, especially for small and medium-sized enterprises (Das, Wang, & Lei, 2006). The micro-payment m-commerce system, which is capable if executing transactions from external online merchants, includes vending machines, tickets, gasoline, and tax fares. In other words, the mobile phone is used as an ATM card or debit card. Time-sensitive and simple procedure transactions are the key success factors to this operation mode.

The mobile phone has also become a new personal entertainment medium. A wide range of interactive entertainment services are available which consist of playing online games, downloading ring tones, watching football video clips, watching live TV broadcasts, downloading music, and so on. According to *Screen Digest* estimates, Korea and Japan accounted for 80% of worldwide games download revenues of Euro 380 million (Screen Digest, 2005). Unsurprisingly, adult mobile services and mobile gambling services are among the fast-growing services. According to Juniper Research, the total revenue from adult mobile services and mobile gambling services could be worth US$1billion and US$15billion respectively by 2008 (Kowk, 2004). Law regulators have to stay ahead of the fast-growing development.

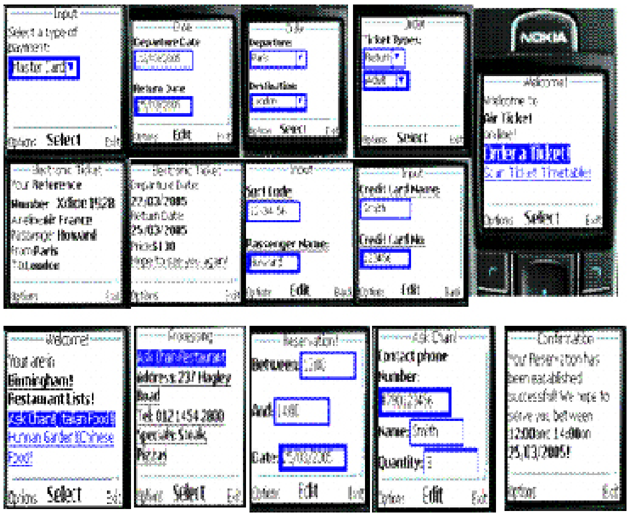
Community tools also generate a large amount of revenue. It evolves from voice and SMS messaging service in the early stage, to the current messenger chatting tools and the distribution of broadband multimedia messaging. Messaging and chatting allow a mobile user to keep contact with the others while he or she is on the move. M-commerce is one of the most important means to communicate in the society.

With different operation modes, each of these can be further classified by the bandwidth utilization (Cherry, 2004):

- **Higher Interactive Multimedia:** Data rate lower than 144kb/s.
- **Narrowband (NB):** Designed as the applications with data rates in the range [144, 384] kb/s.
- **Wideband (WB):** With data rates in the range [384, 2048] kb/s.
- **Broadband (BB):** With data rates higher than 2Mb/s.

As can be seen from this classification, the broadband class is available only when WLAN and other wireless technology access is possible. The NB and WB classes can be distinctive by circuit-switched and packet-switched, therefore the clear evolution path can be drawn as follows: beginning from circuit-switched NB services like basic voice service gradually to the packet-switched WB services then towards to the purely packet-switched BB services.

Figure 2. M-commerce mode prototypes (designed by authors using NMIT 4.0)



RESEARCH FINDINGS

In this research, the Nokia WAP emulator version 4.0 is used to develop the m-commerce application scenarios. Two m-commerce execution scenarios are designed for prototype. The first prototype is concerned with an LBS (location-based service), which provides a list of restaurants that are located near the current location and match a set of user preferences. The second prototype presents a whole financial transaction procedure by purchasing a mobile air ticket. It is more complicated than the first scenario, involving money transaction and payment procedure. Figure 2 shows the entire procedure. The previous steps are similar for both scenarios, which input a set of user preferences, such as departure time, destination, and ticket type. Following this information, the user comes to a secure domain, which can be a financial institution or bank. This step is a significant part in m-commerce applications. The money transaction will be performed in this secure channel by selecting the payment type.

This kind of mobile ticket service creates an extra purchase possibility for public transportation tickets via the mobile phone. Even though the scenarios described above are complicated, from mobile users' point of view, it is transparent and the benefit for them is purchasing goods and request services at anytime, anywhere without constraint of opening hour and physical distribution points, and most importantly it is a cashless payment (Wang, Song, Lei, & Sheriff, 2005).

From the procedures we presented in the two scenarios, it is obvious that there are two critical procedures urgently needing to be solved: user input usability between the client and server, and credit card payment security as performed in financial institutions. As the mobile phone user scrolls the information categories available to be requested and selects the category by pressing a key on the phone pad, a wireless device is dramatically easier to use such that the usability seems to be the critical limitation, one that the user is anxious to solve. And the security relative to a money transaction is still the main concern of business to adapt m-commerce for its intranet and extranet applications.

CONCLUSION

As mobile and wireless technologies are evolving rapidly and sophisticated mobile devices becomes affordable, m-commerce will become a part of our daily lives. The mobile Internet is ideal for particular applications and has useful characteristics that offer a range of services and contents. The widespread adoption of m-commerce is fast approaching.

REFERENCES

BBC News. (2004a). Retrieved March 8, 2004, from http://bbc.co.uk

BBC News. (2004b). Retrieved December 6, 2004, from http://bbc.co.uk

Buvat, J. (2005). Two disruptive technologies. *Land Mobile, 12*(4), 20-21.

Cherry, S. M. (2004). WiMax and Wi-Fi: Separate and unequal. *IEEE Spectrum,* (March).

Das, R., Wang, J. J., & Lei, P. (2006). A social-cultural analysis of the present and the future of the m-commerce industry. In B. Unhelkar (Ed.), *Handbook of research in mobile business: Technical, methodological and social perspective.* Hershey, PA: Idea Group Reference.

Frank, P.C. (2001). *Wireless Web, a manager's guide* (1st ed., pp. 115-132). Boston: Addison-Wesley.

Garber, L. (2002). Will 3G really be the next big wireless technology? *IEEE Computer, 35*(1), 26-32.

Juniper Research. (2006). Retrieved April 20, 2006, from http://www.epaynews.com/statistics/mcommstats.html#49

Kwok, B. (2004). Watershed year for mobile phones. *Companies and Finance in South China Morning Post,* (January 3).

Lamont, D. (2001). *Conquering the wireless world: The age of m-commerce.* New York: Capstone/John Wiley & Sons.

NTT DoCoMo. (2004). *iMode, an overview: Mobile communication and mobile computing.* Retrieved from http://www.rn.inf.tu-dresden.de/scripts_lsrn/Lehre/mobile/print_en/18_en.pdf

Paavilainen, J. (2001). *Mobile business strategies: Understanding the technologies and opportunities* (pp. 32-79). London: Wireless Press.

PriceWaterhouseCoopers. (2001). *2001 global forest & paper industry survey.*

Screen Digest. (2005, February 9). *Mobile gaming gets its skates on.* Retrieved from http://www.theregister.com/2005/02/09/mobile_gaming_analysis

3G Newsletter. (2005). Retrieved January 4, 2005, from http://www.3g.co.uk/PR/Jan2005/8904.htm

Tsalgatidou, A., & Veijalainen, J. (2000, September). Mobile electronic commerce: Emerging issues. *Proceedings of the 1st International Conference on E-commerce and Web Technologies (EC-WEB 2000)* (pp. 477-486), London. Berlin: Springer-Verlag (LNCS 1875).

UMTS Forum. (2003). *Mobile evolution shaping the future.* A UMTS forum white paper.

Wang, J. J., Song, Z., Lei, P., & Sheriff, R. E. (2005, October 3-5). Design and evaluation of m-commerce applications. *Proceedings of the 2005 Asia-Pacific Conference on Communications,* Perth, Australia.

KEY TERMS

Application: An application program (sometimes shortened to application) is any program designed to perform a specific function directly for the user or, in some cases, for another application program. For example, software for project management, issue tracking, file sharing, and so forth.

Bandwidth: A measure of frequency range, measured in hertz, of a function of a frequency variable. Bandwidth is a central concept in many fields, including information theory, radio communications, signal processing, and spectroscopy. Bandwidth also refers to data rates when communicating over certain media or devices. Bandwidth is a key concept in many applications.

Micro-Payment: Means for transferring money in situations where collecting money with the usual payment systems is impractical, or very expensive, in terms of the amount of money being collected.

Mobile Commerce (M-Commerce): The buying and selling of goods and services through wireless handheld devices such as cellular telephones and personal digital assistants (PDAs). Known as next-generation e-commerce, m-commerce enables users to access the Internet without needing to find a place to plug in.

Mobile Internet: Internet access over wireless devices.

Mobile Commerce Adoption Barriers

Pruthikrai Mahatanankoon
Illinois State University, USA

Juan Garcia
Illinois State University, USA

INTRODUCTION

Mobile commerce (m-commerce) emerged as one of the technologies that could change the way consumers engage in electronic business. Consumers have envisioned it as the mobile "electronic commerce," which allows them to purchase goods and services using their wireless mobile devices anywhere, anytime. This mobility, supported by a mobile telecommunications infrastructure, is the major characteristic that differentiates mobile computing from other forms of information technology applications.

Although the widespread use of mobile commerce has been intermingled with advanced telecommunications infrastructure, perceived benefits, and consumer demands, the industry is continuously searching for new and innovative mobile applications. Many consumers are still reluctant to make use of various mobile commerce applications. Technological hype and unreal consumer expectations have generated high hopes for innovative mobile applications that cannot be conceptualized during their initial stages. In many cases, unfilled gaps exist between the potential applications and the actual services provided by leading mobile carriers.

The purpose of this article is to identify and explain different socio-psychological drivers and barriers affecting consumers' motivations to use mobile commerce applications. These determinants are based on our literature reviews and exploratory consumer-based research. We later suggest a research framework to which researchers and practitioners can refer.

BACKGROUND: CONSUMER-BASED DRIVERS OF MOBILE COMMERCE

Mobile computing has two major characteristics that differentiate it from other forms of computing: mobility and broad reach (Turban, Rainer, & Potter, 2006). These two characteristics have created several value-added attributes that drive the demands for mobile-based computing, such as convenience, instant connectivity, and personalization. Wen and Mahatanankoon (2004) capture these demands through their 'aspects of mobility' concept, suggesting that the main driving forces of mobile applications are based on consumers' perception that: (1) their mobile devices are 'always on'; (2) they have the ability to customize their usage according to their lifestyle and social-psychological needs; (3) their location-based services (LBSs) can recognize where they are and then personalize the available services accordingly; and (4) their mobile devices have built-in authentication procedures that support secure mobile transactions. These aspects of mobility have tremendous impact on how consumers perceive various mobile applications.

The success of mobile commerce relies on the synergy of technology innovation, evolution of new value chains, and active customer demand (Zhang, Yuan, & Archer, 2003). These interrelated factors shift the telecommunications industry from being the provider of products or services to being the facilitator of customers' socio-psychological needs. Some practitioners suggest a consumer-centric approach to design effective mobile portals (Chen, Zhang, & Zhou, 2005). A good mobile application not only needs to be ergonomically easy to use, but it also has to provide consumers with sufficient, relevant, and personalized information. The industry should exploit these demand drivers and strengthen them by creating unique sets of innovative mobile applications that interact seamlessly between consumers and their surroundings. To ensure critical mass of mobile commerce adoption, we suggest further development of these existing applications and services to support consumer socio-psychological needs.

Integrated Mobile Devices

These devices are evolving from being a simple telephone with some extra features to an integration of the functionality of a personal digital assistant (PDA) with cellular telephones. The result of such integration creates a device that is able not only to connect to wireless networks, but also to manage organizer features. Speech recognition has become increasingly popular to support mobile commerce activities and will change the nature of user interface design (Fan, Saliba, Kendall, & Newmarch, 2005). In the near future, consumer perspective will change as these technologies are integrated into miniature wearable devices.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Ultramodern Mobile Applications

New and innovative services should exceed today's conventional usage of mobile devices. Enticing future applications will not only blur the boundary between work and play, but will also permit various ubiquitous services to take place simultaneously based on consumer demands (Varshney & Vetter, 2002). Ubiquitous services should be built based on the social network of mobile users.

Geographic-Oriented Applications

Location-based services will be the fastest growing enabler of mobile commerce applications. To consumers, the idea of conducting commercial transactions based on their current location is very appealing (e.g., consumers receiving a coupon for their favorite drink while walking past the coffee shop). In the foreseeable future, various industry consortiums will seek new ways to improve consumers' satisfaction by mapping their usage behaviors to specific locations, times, and events while providing them with options to customize their experience.

Advance Security and Privacy Applications

Two major factors exist concerning security issues: network security and storage security (i.e., securing the information stored in the mobile device). A mobile network needs to protect its users by continuously authenticating its subscribers (Patiyoot & Shepherd, 1999). Biometrics-ready phones can identify and enable authorized users to access the devices' full capabilities while preventing malicious individuals from accessing important personal information. Consumers can also locate lost mobile devices via location-based services by using the embedded global positioning system (GPS) capability.

However, most consumers are not totally convinced that mobile commerce would be a satisfactory experience. Consumers think twice before engaging in mobile commerce since most mobile commerce functionalities and services are not similar to those of electronic commerce. Mobile commerce is not intended to replace electronic commerce, but rather supplement it. Various factors, such as user interface, network speed, and users' self-efficacy, hinder many potential mobile applications.

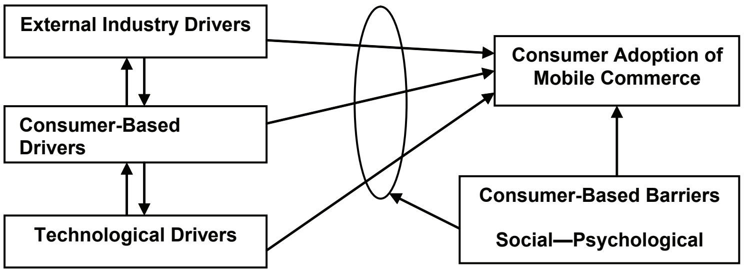
BARRIERS TO MOBILE COMMERCE ADOPTION

Based on our preliminary findings, we are able to identify six main consumer-based barriers to mobile commerce adoption. These socio-psychological barriers are tightly integrated and include unawareness, device inefficiency, personalization/customization, nice-to-have/must-have, roaming, and electronic commerce perception. A successful solution to these interrelated factors will most likely result in mobile commerce reaching its critical mass. Figure 1 suggests a research framework on mobile commerce applications. These integrated barriers directly impact the mobile commerce adoption, as well as moderate the strength of external industry, technological, and consumer-based drivers. The external industry and technological drivers directly influence the consumer-based drivers and vice versa.

Unawareness Barrier

Awareness of mobile commerce existence implies that the individual has heard of it and has some idea of the kind of services it provides. Consumers are not always aware of their wireless devices' mobile commerce capabilities or their carrier's pricing scheme. Sometimes mobile carriers fail to communicate the mobile commerce capabilities to consumers. Only a few active users explore their mobile devices beyond voice communications and information-seeking activities. With various third-party electronic commerce vendors joining the bandwagon, it is often up to the users to discover how to connect to the Internet, download applications, or figure out how to use such applications. Therefore, in many mobile usage settings, consumer self-efficacy generally plays a significant role in exploring ground-breaking functionalities.

Figure 1. Consumer-based mobile commerce adoption framework



Device Inefficiency Barrier

The inefficiency of small mobile devices continues to be a problem. Every extra navigational input reduces the possibility of a transaction by 50% (Clarke, 2001). In addition to the limitations of screen size, power, and processing capability, device manufacturers need to be aware that consumers have a variety of multi-tasking activities (Lee & Benbasat, 2003). Personalization can compensate for the drawbacks of a small user interface (Ho & Kwok, 2003), but it may negatively affect other aspects, such as privacy and security. In many aspects, mobile user interfaces need to be designed to support users' limited but ever-shifting tasks.

Personalization/Customization Barrier

Customization services are context-specific applications that target each individual. These operations range from customized ring tones to location-based services. Since most customizable systems typically store users' essential information, issues related to privacy will be a major concern for consumers. These concerns for individual privacy negatively impact the adoption of mobile commerce. Consumers fear that they can be profiled, and their purchase history and navigation behaviors analyzed and abused (Pitkow et al., 2002). The lack of trust also leads to consumer avoidance of personalization/customization mobile applications.

Nice-to-Have vs. Must-Have Attitudinal Barrier

The industry must move beyond nice-to-have services and devise new 'must-have' services that positively affect people's lives (Jarvenpaa, Lang, Takeda, & Tuunainen, 2003). A nice-to-have feature may tempt consumers into buying a mobile device, but it cannot sustain a steady stream of revenue for the industry. Mobile application developers are searching for their killer application without trying to understand the socio-psychological aspects of hedonic mobile usage activities. Despite more than 40 inventive mobile applications, only five consumer-based applications are considered must-have applications (Mahatanankoon, Wen, & Lim, 2005); these are location-based, banking, entertainment, Internet, and emergency applications. Mobile designers must take into consideration usage environments that are relatively unstable and dynamic, potentially changing in a matter of seconds (Tarasewich, 2003).

Geographical Roaming Barrier

Consumers should be able to use the same mobile devices and services from anywhere in the world. Interoperability relates to the ability to use the same mobile device anywhere in the world. However, interoperability that accrues significant charges hinders the rapid adoption of mobile applications. Due to the competitive nature of the telecommunications industry, third-party providers and mobile carriers generally design their mobile applications based on device characteristics and specific network standards, which do not support communication and information sharing across mobile device platforms. Open Mobile Alliance (OMA) and the World Wide Web Consortium (W3C) have set their goal to create a global and interoperable mobile commerce market. It is hopeful that these consortiums will more closely connect worldwide consumers.

Perceptual Barrier

Many characteristics of traditional and electronic commerce can impact the way consumers perceive mobile commerce. Prior exposure to electronic commerce applications can have a significant impact on consumers' tendencies to modify their behaviors to fit the nature of small handheld devices (Orlikowski & Gash, 1994). Trust and trustworthiness of mobile commerce are still questionable. The idea of not dealing with somebody face to face at a physical location, or not being able to touch the merchandise, may sound unattractive to consumers. Many users simply do not like the idea of entering personal information and credit card numbers into their mobile transactions, fearing unsecured wireless networks or becoming potential victims of identity theft when the devices are stolen. Siau and Shen (2003) recommend that building customer trust in mobile commerce is a continuous process. Nevertheless, unlike electronic commerce's virtual communities, mobile commerce still lacks the sense of *virtualness* among consumers (e.g., customers cannot interact with other customers and gain feedback about a merchant from other customers).

These foremost barriers suggest that the mobile commerce buying experience is totally different than the traditional or electronic commerce buying experience. Electronic commerce customers may decide to buy products from a trusted vendor just by looking at its reliability and reviews, but for mobile commerce consumers, this functionality still remains a challenge. The industry is obligated to assist customers to overcome such barriers before it can reap any potential revenue from mobile commerce.

FUTURE TRENDS OF MOBILITY RESEARCH

Enormous potential exists for mobile commerce applications in the future, although the industry is still searching for distinct and profitable business models. Its current hype has surpassed its usefulness, but the joint efforts of industry players are pushing mobile commerce to become a widely used consumer-based technology. Practitioners and researchers can examine these drivers/barriers and their impact on consumers' socio-psychological behaviors. As mobile commerce evolves and sets its goal on changing the way consumers interact with the world, it is necessary to explore and take into account the obstacles that prevent the technology from reaching its true potential.

CONCLUSION

This article discusses the most salient characteristics of mobile commerce based on its mobility and success attributes. These interrelated factors help identify different consumer-based drivers of mobile commerce adoption, such as ultramodern, geographical-oriented, and advance security/privacy applications. The article then discusses the socio-psychological barriers of mobile commerce adoption. Various factors such as device inefficacy, interoperability, users' perceptions, and self-efficacy hinder many potential mobile applications. Given these limitations, mobile commerce applications should exploit the demand drivers and strengthen them by creating their own set of unique and innovative mobile applications.

REFERENCES

Clarke, I., III. (2001). Emerging value propositions from mobile commerce. *Journal of Business Strategies, 18*(2), 133-148.

Fan, Y., Saliba, A., Kendall, E. A., & Newmarch, J. (2005) Speech interface: An enhancer to the acceptance of mobile commerce applications. *Proceedings of the International Conference on Mobile Business.* Los Alamitos, CA: IEEE Computer Society Press.

Jarvenpaa, S. L., Lang, K. L., Takeda, Y., & Tuunainen, V. K. (2003). Mobile commerce at a crossroads. *Communications of the ACM, 46*(12), 41-44.

Lee, Y. E., & Benbasat, I. (2003). Interface design for mobile commerce. *Communications of the ACM, 46*(12), 49-52.

Mahatanankoon, P., Wen, H. J., & Lim, B. (2005). Consumer-based m-commerce: Exploring consumer perception of mobile applications. *Computer Standards and Interfaces, 27*(4), 347-357.

Orlikowski, W. J., & Gash, D. (1994). Technological frames: Making sense of information technology in organizations. *ACM Transactions on Information Systems, 12*(2), 174-207.

Patiyoot, D., & Shepherd, S. J. (1999). Cryptographic security techniques for wireless networks. *ACM SIGOPS Operating Systems Review, 33*(2), 36-50.

Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., et al. (2002). Personalized search. *Communications of the ACM, 45*(9), 50-55.

Siau, K., & Shen, Z., (2003). Building consumer trust in mobile commerce. *Communications of the ACM, 46*(4), 91-94.

Tarasewich, P. (2003). Designing mobile commerce applications. *Communications of the ACM, 46*(12), 57-60.

Turban, E., Rainer, K., & Potter, R. (2006). *Introduction to information technology* (3rd ed.). New York: John Wiley & Sons.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*(3), 185-198.

Wen, H., & Mahatanankoon, P. (2004). Mobile commerce operation modes and applications. *International Journal of Electronic Business, 2*(3), 301-315.

Zhang, J., Yuan, Y., & Archer, N. (2003). Driving forces for mobile commerce success. In M.J. Shaw (Ed.), *E-business management: Integration of Web technologies with business models* (pp. 51-76). Boston: Kluwer Academic.

KEY TERMS

Global Positioning System (GPS): A satellite-based tracking system that enables the determination of a GPS device's location.

Location-Based Service (LBS): One of several mobile services and applications offered to consumers via the utilization of GPS technology via the mapping of existing spatial information.

Open Mobile Alliance (OMA): An alliance of leading mobile operators, device and network suppliers, information technology companies, and content providers to create a universal mobile interoperability standard.

Personal Digital Assistant (PDA): A small handheld organizer, sometimes equipped with operating systems and wireless Internet capability.

Mobile Commerce (M-Commerce): The process of conducting electronic commerce activities through small mobile devices, such as mobile phones, pocket PCs, or PDAs.

Smart Phone: Internet-enabled cell phone that can support mobile applications and generally having advanced microprocessors to support various mobile applications.

Socio-Psychology (Social Psychology): A study of psychology related to the behaviors of groups and the influence of social factors on an individual.

Wireless Mobile Computing: The combination of mobile devices used in a wireless environment.

A Mobile Computing and Commerce Framework

Stephanie Teufel
University of Fribourg, Switzerland

Patrick S. Merten
University of Fribourg, Switzerland

Martin Steinert
University of Fribourg, Switzerland

INTRODUCTION

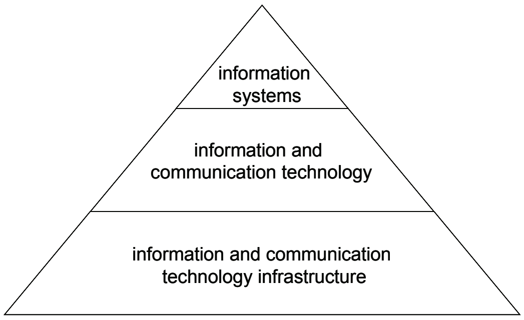
This encyclopedia on mobile computing and commerce spans the entire nexus from mobile technology over commerce to applications and end devices. Due to the complexity of the topic, this chapter provides a structured approach to understand the interrelationship in-between the mobile computing and commerce environment. A framework will be introduced; the approach is based on the Fribourg ICT Management Framework, elaborated at our institute with input from academics and practitioners, which has been tried and tested in papers, books, and lectures on ICT management methods. For published examples, please consult Teufel (2001, 2004), Steinert and Teufel (2002, 2004), or Teufel, Götte, and Steinert (2004).

THE MOBILE CONVERGENCE CHALLENGE

The information revolution has drastically reshaped global society and is pushing the world ever more towards the information-based economy. In this, information has become a commodity good for companies and customers. From an economical perspective, the demand for information at the right time and place, for the right person, and with minimal costs has risen. The transformation towards this information-driven society and economy is based on the developments of modern *information and communication technology (ICT)*. Different industries are able to generate enormous synergy effects from the use of ICT and the *information systems (IS)* building on these technologies, especially the Internet. It is a possible instrument to chance the structure and processes of entire markets.

As shown in Figure 1, information and communication technology can be differentiated in its infrastructure, the technologies themselves, and the information systems running on these technologies. In general, the infrastructure consists of all hardware- and software-related aspects as well as human resources. Consequently, the technologies themselves enable the collection, storage, administration, and communication of all data. These data can be used to synthesize information in respective systems, supporting the decision process and enabling computer-supported cooperative work.

Figure 1. Information and communication technology, infrastructure, and systems


information systems
information and communication technology
information and communication technology infrastructure

The term information and communication technology (ICT) appeared in recent years. Due to the harmonization of *information technology (IT)* and the digitalization of the *telecommunications (CT)* infrastructure and the liberalization of the latter business sector, the ICT market established itself (see Figure 3). Consequently, the development and convergence of ICT became increasingly complex. Figure 2 illustrates the associated technology convergence.

Nowadays, a new aspect has entered the arena: mobility. Mobility is perhaps the most important trend on the ICT market. The fundamental characteristic of mobile technologies is the use of the radio frequency band for (data) communication, which is often referred to as "wireless." The "wireless trend" has influenced not only the telecommunications and IT sector, but also most traditional markets, in the same way wired ICT did before. In addition, a convergence of wired and wireless, respectively fixed and mobile ICT can be observed.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Technology convergence (Teufel, 2004, p. 17)

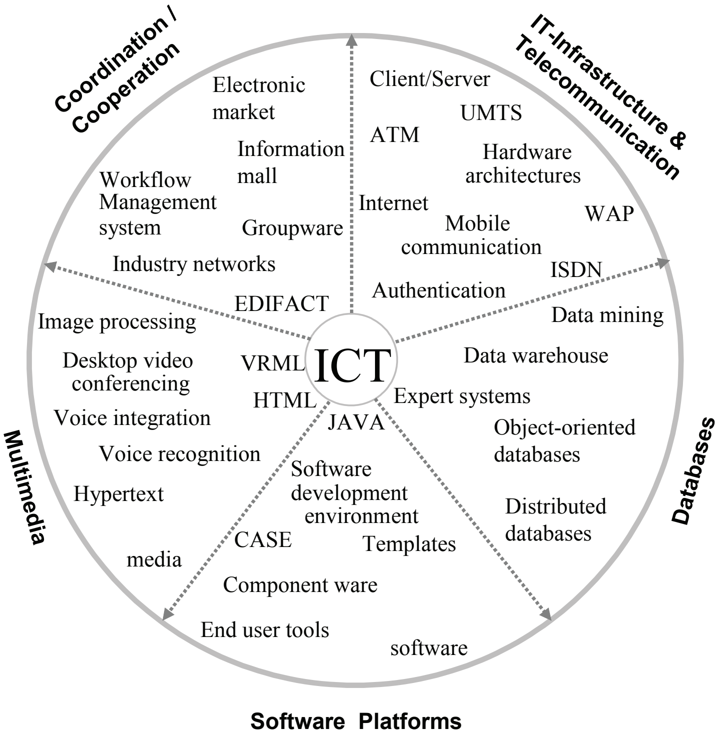

Coordination / Cooperation
IT-Infrastructure & Telecommunication
Databases
Software Platforms
Multimedia
ICT
Electronic market
Information mall
Workflow Management system
Groupware
Industry networks
EDIFACT
Client/Server
UMTS
ATM
Hardware architectures
Internet
Mobile communication
WAP
ISDN
Authentication
Data mining
Data warehouse
Expert systems
Object-oriented databases
Distributed databases
Templates
software
Software development environment
CASE
Component ware
End user tools
JAVA
HTML
VRML
Image processing
Desktop video conferencing
Voice integration
Voice recognition
Hypertext
media

Figure 3. Mobile and fixed-line ICT convergence

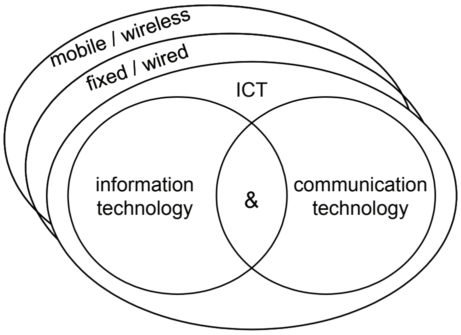
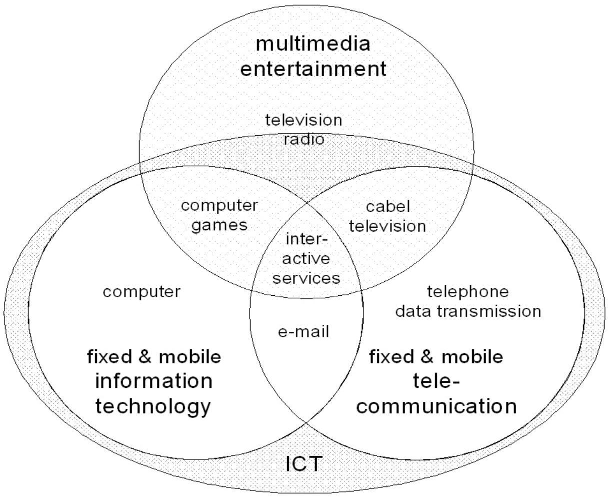

mobile / wireless
fixed / wired
ICT
information technology
&
communication technology

Figure 4. ICT and multimedia entertainment convergence (Teufel, 2004, p. 14)


multimedia entertainment
television
radio
computer games
cabel television
inter-active services
computer
telephone
data transmission
e-mail
fixed & mobile information technology
fixed & mobile tele-communication
ICT

As shown in Figure 3, the convergence of information technology and communication technology to ICT can be seen as the first phase of convergence. This was caused by the digitalization and liberalization in the telecommunications sector. The next phase of convergence was the success of mobile ICT, initializing a competition between wireless and fixed ICT. Meanwhile, information and communication as well as mobile and wired technologies have not only coexisted; they have merged, generating enormous synergy effects for both business and customer. In addition, another not just technological convergence can be observed. The entertainment and multimedia branch has entered the ICT market and vice versa, as illustrated in Figure 4.

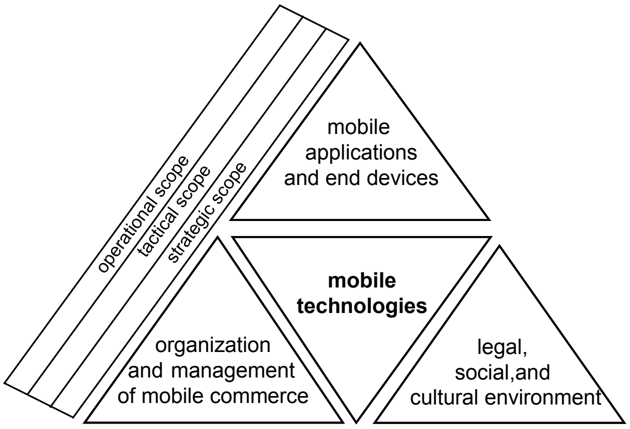
The trend shown in Figure 4 becomes obvious when looking at the boom in interactive games or home cinema computerized equipment—again accelerated by the digitalization in a sector, this time the television (DVB) and radio

(DAB). Again, the Asian market is leading edge. In South Korea, they are already running a fully functional system, based on the digital mobile broadcasting standard (DMB), bringing video broadcasting directly to the mobile end-device via satellite (tu4u, 2006). Finally, the three dimensions, fixed and mobile ICT convergence plus entertainment/multimedia, form the core of this encyclopedia's topic: the challenges of mobile computing and commerce.

THE MOBILE COMPUTING AND COMMERCE FRAMEWORK

Mobile computing and commerce comprises all business processes between administration, business, and customer via public or private wireless communication networks and with value creation. To understand the actual trends, recognizing the possibilities and threats and coping with the challenges of mobile computing and commerce are complex tasks. It becomes obvious that mobile computing and commerce consists of multiple dimensions, which are, in addition, interrelated. In order to structure the discussion, a framework for mobile computing and commerce is introduced. Using the classical scientific engineering approach, the framework allows a detailed analysis of single aspects and a reintegration of the diverse solutions in the synthesis. Furthermore, it covers the main issues, controversies, and problems from a market and business perception. Figure 5 features this ,mobile computing and commerce framework.

Figure 5. Mobile computing and commerce framework


operational scope
tactical scope
strategic scope
mobile applications and end devices
mobile technologies
organization and management of mobile commerce
legal, social,and cultural environment

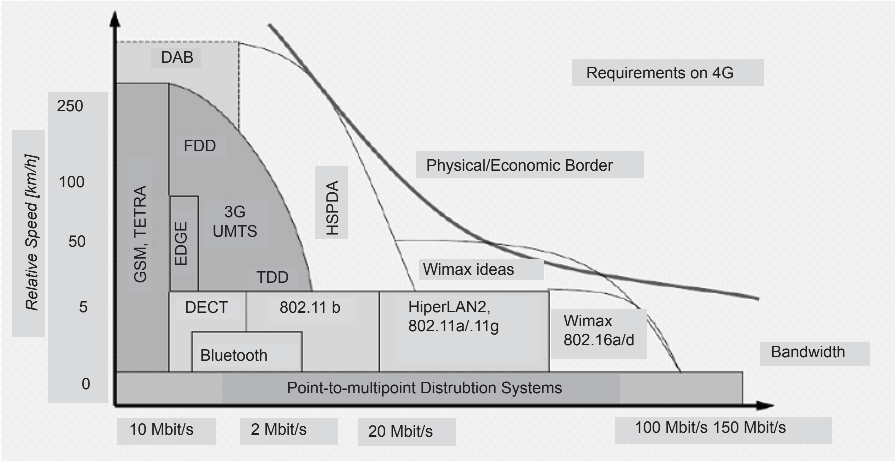
The four different dimensions of the framework as demonstrated in Figure 5 in addition show a common underlying scope. The strategic scope covers issues of long-term influence (more than five years of impact), as the tactical scope deals with all aspects in a timeframe of one to five years. Finally, all short-term topics are subject of the operational scope and handled within a year's period. The individual four main parts of the framework are examined in the following sections.

Mobile Technologies

The origin and foundation of every case of mobile computing and commerce are mobile technologies. They are the centerpiece of the framework and comprise the different technological aspects. They are building the foundation for discussing all other aspects of the framework. Mobile technologies have evolved rapidly in the last decade, not only gaining market penetration, but in terms of bandwidth and relative speed. Figure 6 presents today's available wireless access technologies—also introducing a physical and economic border.

As such, this dimension includes aspects which are dealt with in the categories mobile information systems, mobile service technologies, and enabling technologies of this encyclopedia.

Figure 6. Wireless access technologies (adapted from Schiller, 2003, p. 450)



Legal, Social, and Cultural Environment

In a mobile environment, corporate social responsibility (CSR) is a fairly new field of increasing attention. It deals with the consequences of globalization, economic and ecological disaster, as well as financial affairs and others. Referring to the Global Compact Program 2000 from the United Nations and the Green Paper on CSR from the European Union, principles and guidelines are available today. These have led to programs that enable companies to continuously analyze and handle the versatile influences and effects on society and vice versa (Teufel et al., 2004).

Furthermore, the existence, use, and diffusion of mobile technologies are also strongly influenced by environmental aspects, especially from legal, social, and culture sub-environments. Examples are data protection issues, surveillance discussions, and radiation concerns, respectively. Other issues may also include important aspects such as standardization and regulation.

Furthermore, mobile technologies also change the way of living—introducing new concepts like mobile working. Especially the new work-life-(un)balance is subject to heated debates. Mobile technologies, applications, and end devices not only represent new opportunities in a business environment, but also create an interconnected and virtual world. In this, the digital divide more and more becomes a critical threat. Therefore topics of the categories like "mobile enterprise implications for society, business, and security" are to be considered.

Organization and Management of Mobile Commerce

To cope with the business challenges of mobile computing and commerce, all company internal aspects of the organization and the management of mobile commerce form a particular topic space. First of all, the classical roles of the CIO and the CTO have to be re-evaluated, taking mobile ICT into account. Furthermore, mobility also affects a whole set of management issues, which have already been previously influenced by fixed ICT. For example, the information management has to consider the aspect of mobile working when planning information system architectures. This in turn results in an adoption of current business processes and workflow implementations. Especially the procurement and distribution processes go through a fundamental change. In addition, mobile ICT offers new possibilities in the customer relationship management.

This dimension for example includes aspects described in "Mobile Commerce and E-Business."

Mobile Applications and End Devices

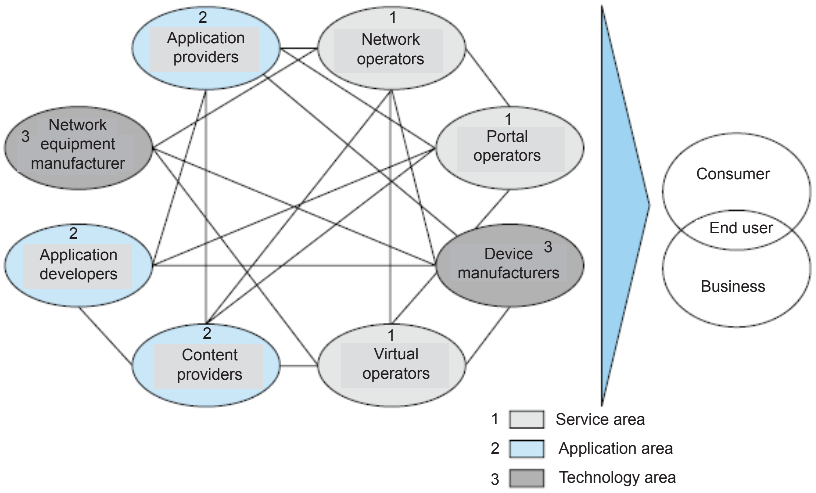
The focus point of every examination of mobile computing and commerce is the actual applications and end devices it is running onto. Again the Asian market can be consulted, to give an example of cutting-edge end device research. NTT DoCoMo is working on a future mobile phone device that uses human fingers as receiver. For this, a wristwatch-like bone conduction terminal is used in contact with the human arm (NTT DoCoMo, 2006). Above all, the Asian market is leading the way towards an all IP-based mobile network environment. Thus this last but probably most important framework dimension features topics such as: mobile to "consumer applications", "mobile applications for the extended enterprise", and "enabling applications."

CONCLUSION

Throughout the previous sections, it has been shown that coping with the challenges of mobile computing and commerce is a complex problem. Therefore, and to structure this encyclopedia, the framework has been introduced. It aims to provide managers, engineers, and practitioners with a profound approach to handle fixed and mobile information and communication technology. In such a mobile computing and commerce environment, the different market players themselves can be subsequently differentiated as shown in Figure 7, following the EITO on their special on "Entering the UMTS era—mobile applications for pocket devices and services."

Figure 7 illustrates the mobile data value net. In this interconnected environment, the nine different market players experience an enforced competition, due to the fact that every action influences the entire business network. As a result, a duality of competitive and cooperative business strategies established itself (Steinert & Bult, 2004, p. 31) to generate network effects, introducing the phenomenon of co-opetition (Brandenburger & Nalebuff, 1996). On a more general level, this again shows the complexity of a mobile computing and commerce environment.

Figure 7. Mobile data value net (EITO, 2002, p. 205)



REFERENCES

Brandenburger, A., & Nalebuff, B. (1996). *Co-opetition* (1st ed.). New York: Doubleday.

EITO (European Information and Technology Observatory). (2002). *Eito report 2002.*

NTT DoCoMo. (2006). *R&D.* Retrieved January 12, 2006, from http://www.nttdocomo.com/corebiz/rd/index.html

Schiller, J. (2003). *Mobile communication* (2nd ed.). London: Addison-Wesley.

Steinert, M., & Bult, A. (2004). Strategische unternehmensführung von hightech-unternehmen—insights von swisscom-fixnet. In S. Teufel, S. Götte, & M. Steinert (Eds.), *Managementmethoden für ICT-unternehmen* (p. 12).

Steinert, M., & Teufel, S. (2002). The Asian lesson for mobile provider—An all-out strategic paradigm shift. *Proceedings of ITU Telecom Asia 2002* (pp. 25-44), Hong Kong.

Steinert, M., & Teufel, S. (2004, September 17-19). Beyond e-business—why e-commerce and Web organizations should monitor the mobile dimension. *Proceedings of the 2nd International Conference on Knowledge Economy and Development of Science and Technology (KEST2004)* (pp. 446-454), Beijing, China.

Teufel, S. (2001, August 6-12). ICT-management framework. *Proceedings of the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR 2001)* (pp. 9-24), L'Aquila, Italy.

Teufel, S. (2004). Managementmethoden für ICT-unternehmen—dargestellt mittels dem Fribourg ICT management framework. In S. Teufel, S. Götte, & M. Steinert (Eds.), *Managementmethoden für ICT-unternehmen.* Zurich: Verlag Industrielle Organisation/Orell Füssli.

Teufel, S., Götte, S., & Steinert, M. (Eds.). (2004). *Managementmethoden für ICT-unternehmen: Aktuelles wissen von forschenden des iimt der Université Fribourg und spezialisten aus der praxis.* Zürich: Verlag. Industrielle Organisation.

tu4u. (2006). *TU media corporation.* Retrieved January 12, 2006, from http://www.tu4u.com/

KEY TERMS

Co-Opetition: Following Brandenburger and Nalebuff (1996), co-opetition is the economic situation between a company and a competing company that provides complementary products and services. Following game theory, a differentiated approach strategic than the generic competitive strategies are necessary (see also *ValueNet*).

Fribourg ICT Management Framework: The framework has been elaborated at the International Institute of Management in Technology (IIMT) of the University of Fribourg (Switzerland) with input of academics and practitioners. It provides an integrated approach to cope with the business challenges of the information-based economy.

Information and Communication Technology (ICT): The result of developments in the fields information technology (IT) and communication technology (CT), and their convergence caused by the digitalization and liberalization in the telecommunication sector.

Legal, Social, and Cultural Environment: This framework dimension covers all aspects and implications of Mobile ICT for Society and Business.

Mobile Application and End Device: Mobile applications running on mobile end devices are the topic of this framework dimension.

Mobile Computing and Commerce Framework: The framework is based upon the Fribourg ICT Management Framework and presents an integrated view on the different fields to be considered, while examining the issues and controversies of mobile computing and commerce.

Mobile ICT Convergence: As ICT can be seen as the first phase of convergence, mobile ICT convergences introduce wireless technologies next to wired ICT.

Mobile Technology: Wireless mobile access technology and the centerpiece of the framework.

Network Effect: Following Katz and Shapiro (1985), each new network participant directly increases the benefit of all other actors in a network, for example, by offering a new communication possibility (primary or direct network effect); an increased size of a network also indirectly increased the value of the entire network indirectly, for example by pushing an industry standard (secondary or indirect network effect).

Organization and Management of Mobile Commerce: All company internal aspects of the organization and the management issues, which are influenced by mobile ICT, also including aspects such as mobile business.

ValueNet or ValueWeb: Instead of a linear value chain, the company, its suppliers, and customers, and also its complementors and competitors, form a ValueNet or ValueWeb. Co-opetition, reciprocal actions, and network effects must be taken into account in the economics of such a value net.

Mobile E-Commerce as a Strategic Imperative for the New Economy

Mahesh S. Raisinghani
TWU School of Management, USA

INTRODUCTION

A new form of technology is changing the way commerce is being done globally. This article provides an overall description of mobile commerce and examines ways in which the Internet will be changing. It explains the requirements for operating mobile commerce and the numerous ways of providing this wireless Internet business. While the Internet is already a valuable form of business that has already changed the way the world is doing business, it is about to change again. Telecommunications, the Internet, and mobile computing are merging their technologies to form a new business called *mobile commerce* or the *wireless Internet.* This is being driven by consumer demand for wireless devices and the desire to be connected to information and data available through the Internet. There are many new opportunities that have only begun to be explored, and for many this will become a large revenue source for those who capitalize upon this new form of technology. However, like other capital ventures, these new opportunities have their drawbacks, which may limit growth of the mobile commerce market if not dealt with. Mobile e-commerce technology is changing our world of business just as the Internet alone has changed business today.

BACKGROUND

Mobile commerce is the delivery of electronic commerce capabilities directly into the consumer's hand via wireless technology and putting a retail outlet in the customers' hand anywhere. This form of e-commerce allows businesses to reach consumers directly regardless of their location. The term mobile commerce or m-commerce is a variation of the e-commerce or electronic commerce term used for business being done over the Internet. Known as next-generation technology, m-commerce enables users to access the Internet without the need to find a place to plug in. There are signs that m-commerce is growing in popularity. Gartner Research (2004) forecasts that in six years time, 60% of people aged 15 to 50 in the European Union and the United States will wear an always-on wireless communications device for at least six hours a day, and more than 75% will do so by the year 2010.

Mobile commerce is the integration of technologies using wireless devices for conducting business over the Internet. M-commerce can be done by computer solutions, such as laptops and palm pads, with wireless devices attached to connect to the Internet or by using newly adapted cellular phones to receive digital transmissions of Internet material to these phones. These are all linked by software and service providers which provide the platform to conduct these operations.

A new business model is emerging: the integration of wireless networks with data communications, combined with electronic commerce, to create wireless e-commerce. Wireless e-commerce will generate significant revenues within the next several years from such services as wireless banking, wireless stock trading, and a variety of wireless-based shopping ventures. Wireless communications and e-commerce already are multi-billion-dollar global businesses. The integration of mobile communications with e-commerce has already started. For years companies in the vertical markets, such as field repair, have been utilizing mobile communications networks to enable their technicians to order parts and check inventories. The opportunities for wireless e-commerce in the horizontal markets, such as traveling executives and the consumer markets, is generating much appeal (Reiter, 1999).

M-commerce is a Quantum leap of technology applications and will not be limited simply to banking and brokerages. Other market uses will emerge. Payment options are one example that are being tested now in which products in a store may be scanned as one walks out and automatically deducted from a *smart card* which stores cash on it from your local bank. Airline and rail connections will be enhanced with ticket reservation and payment facilities. Mobile phone users will also have access to new online auction houses to submit bids and check developments by use of the cellular phone (Brokat, 2000).

MOBILE COMMERCE: STATE OF THE INDUSTRY

According to the Strategis Group (2005), by the year 2010, there will be one billion wireless subscribers worldwide on 3G (third-generation) networks. ARC Group estimates that

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Payback period for wireless LAN

	Retail	Manufacturing	Healthcare	Office Automation	Education
Benefits per company (millions $)	5.6	2.2	.94	2.5	.5
Costs per company (millions $)	4.2	1.3	.90	1.3	.3
Payback (# of months)	9.7	7.2	11.4	6.3	7.1

Table 2. Wireless Internet users

Region		2001	2004	2010
USA	Internet Users (#M)	149	186	247
	Wireless Internet Users (#M)	5.5	23	91
	Wireless Internet User Share (%)	3.7	12.5	35.1
Worldwide	Internet Users (#M)	552	941	1,781
	Wireless Internet Users (#M)	79	200	779
	Wireless Internet User Share (%)	14.4	21.2	43.8

by 2007 approximately 546 million users will spend close to $40 billion on mobile commerce (Schone, 2004). The reasons for this phenomenal growth are attributed to business factors such as substantial increase in remote workers and the telecommuters' need for improved customer service; the economic justification of mobile computing solutions through productivity gains and competitive advantages gained by early implementers; availability of inexpensive hardware with pre-packaged vertical industry application solutions, and less expensive and faster wireless networks; convergence of the Internet, wireless, and e-commerce technologies; and emergence of location-specific and mobile commerce applications, especially by a socially upscale and mobile population. As illustrated in Table 1, the data from the Wireless LAN Association (2005) shows that across all industries, with all economic benefits such as increased productivity, organizational efficiency, and extra revenue/profit gain considered, the wireless LAN paid for itself within 12 months time.

By 2008, a tenth of the world's mobile phone users will use their handsets as video players and cameras, and to download news, sports, and entertainment news (Gartner Research, 2004). Table 2 lists the wireless Internet users from 2001 and 2004, and forecasts the statistics for 2010 (eTForecasts, 2006).

Figure 1 summarizes the expected growth in m-commerce revenues over the period of 2001-2006.

International Data Corporation's (IDC's) forecast shows that total mobile data revenues are expected to be increasing by more than 31% per annum, whereby the CAGR for revenues generated by m-commerce is estimated to be more than 265% per annum. As seen in Figure 1, despite rapid growth, income from m-commerce will remain a very small portion of total data revenues (highest value 5.2% in 2006). This implies that in the short run m-commerce will not turn in profits to justify the investments in the new technologies that were initially believed to boost it. This applies to Europe and the United States. Asia, on the other hand, has developed more quickly with respect to m-commerce. Figure 2 illustrates the mobile operator data services revenue from 2001 to 2006 in Hong Kong and China.

The combined figures for Hong Kong and China show that the total mobile data services revenues are expected to increase for the period 2001-2006 by more than 70% per annum. The growth in m-commerce revenues in Hong Kong and China is currently 26.3% and is expected to outstrip 45% by 2006.

This outstanding acceptance of m-commerce in Asia is only partially due to the currently used mobile technologies that require fewer investments for the upgrade to 3G. The major drivers behind this trend are the habits of the Asians who are keener on innovative technologies, and the variety of content providers that attracts an increasing number of mobile services users.

Interestingly in Asia, the investments were restricted to the lower priced 3G licenses. The service providers, therefore, are able to offer their services at a much lower cost. Especially in China and Hong Kong, where mobile technology was introduced at a later stage and the 2G CDMA standard was adopted initially, the transition to 3G required little investments for the upgrade of the networks. This allows them to offer the services at a lower cost compared to the potential

Figure 1. Western Europe mobile operator data services revenue 2001-2006 (Source: IDC, "Mobile Data Platforms and Services in Western Europe Forecast and Analysis 2001-2006")

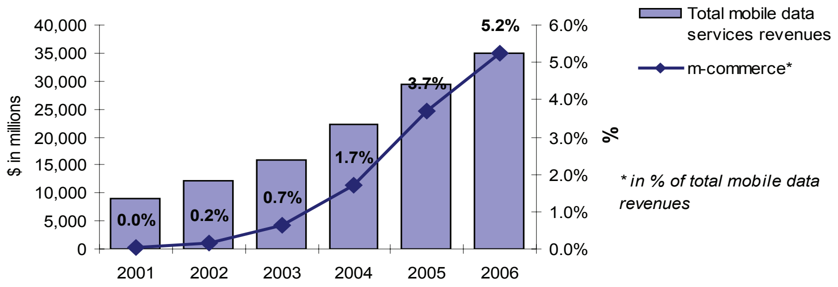
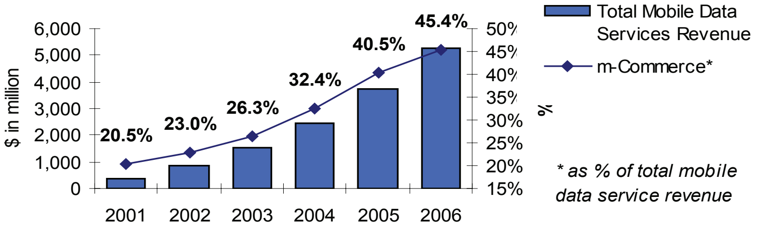


Figure 2. Hong Kong and China mobile operator data services revenue 2001-2006 (Source: IDC, "Asia/Pacific M-Commerce Forecast and Analysis: Opportunities Await")



costs to the European consumers. The lower costs make it easier for consumers to try this new technology, become accustomed to, and ultimately adopt it.

The mobile market penetration in the U.S. is around 41%, far less than countries like Finland with 75%, Hong Kong with 89%, or the United Kingdom with 74% (Magura, 2003). Besides, only 6% of users in the U.S. use their mobile phones to access the Internet, and this is a much lower percentage compared with other countries like Japan with 72%, Germany with 16%, or the United Kingdom with 10% (Beal et al., 2001, p. 6).

CONDITIONS SPARKING DEVELOPMENT OF MOBILE COMMERCE

A large reason for the high level of acceptance for m-commerce is the large number of mobile phone users and Internet users around the world. The Internet has promoted electronic services, and m-commerce is another means of using the Internet. Customers now want to take advantage of Internet services from mobile end devices so that they can conduct business from any location in the world.

The highly lucrative industry of Internet commerce and mobile communication is a driving force in bringing many companies to develop this technology. The second catalyst is that many mobile phone users, especially in Europe and then in South East Asia, will be using smart cards, and this technology is another way to use cellular phones for business. Much of today's wireless e-commerce technology is a result of technology being developed by many of the mobile phone makers. The Europeans have led this charge since they have some of the highest numbers of cellular phone users. This is a result of the global economic and political environment during the 1980s, which promoted greater unification and collaboration, which helped the new telecommunications industries in Europe to flourish. As the need for better communication facilities grew, due to increased trade and investment flows, the solutions provided by the new technological developments become more viable.

EQUIPMENT USED IN MOBILE COMMERCE

This is an overview of the equipment necessary to conduct mobile commerce. It consists of digital cellular phones, smart cards, laptops, or palm pads, and the software to operate and communicate with this hardware.

Digital Cellular Phone

Dual-slot mobile phone technology was developed in Europe. Two key things are required for it to work: phones with the capability and chip-based cards. Dual-slot mobile phones offer a suite of value-added services, including mobile banking. It can be reprogrammed in the field and has substantial free memory for further applications. This allows subscribers to turn their mobile phones into tools to support their business and leisure lifestyle (Brokat, 2000).

To understand digital cellular technology, we must understand the background of the cellular phones as it relates to speed of data transmissions. Analog technology is considered to be the first generation of cellular technologies. The second generation of digital cellular technology is high-speed circuit-switched and packet-switched data technology. High-speed circuit-switched data technology uses a single voice channel and delivers data at a rate of 9.6 Kbps. Packet-switched means the computer that is connected to the cell phone sends and receives bursts, or packets, across the radio channel. The channel is occupied only for the duration of the data transmission instead of continuous transmissions, making it more efficient than circuit switched. The third generation (3G) of digital cellular technology refers to a much higher data transmission speed in the range of 14.4 Mbps. It will enable wireless multimedia applications such as videoconferencing. 3G is the collective term used for several engineering proposals to make wireless networks more data capable than first-generation analog and second-generation digital cellular networks. Some of the challenges of network speed and volume capability are addressed by these networks which must be able to transmit wireless data at 144 kilobits per second at mobile user speeds, 384 kbps at pedestrian user speeds, and 2 megabits per second in fixed locations (Schone, 2004).

Smart Cards

Smart cards are a cross between an ID card and electronic wallet. They can be used to store and exchange money from banks as well as support the payment functions of digital cellular phones. Already used in parts of Europe, this card provides many attributes that will enable technology to better serve consumers. Some present mobile communication in Europe relies on dual smart card technology. It consists of one smart card, internal to the cell phone, and one external card, which can hold personal information and be used as a cash card or electronic wallet, as well as phone card (Rundgren, 1999b).

Smart cards are a plastic ID card containing an integrated circuit chip that is capable of reading, writing, storing, and processing information. The size and shape of the plastic, the positioning of the chip, and its resilience to attack are defined by international standards. They cost between $2 and $20 depending on their capabilities. Multiple applications include contactless smart cards that can be read by radio signal from a card reader.

Smart cards have won the battle with magnetic-strip cards because of their security, reliability, capability, and lifetime cost. A contactless smart card ticketing solution is much cheaper in the long run. Capital investment can be 90% lower, revenues can be increased by 5% to 10% through lower fraud, and maintenance can be 30 times lower with a contactless smart card system. Smart cards are already used for public transport and parking services in cities in Europe and Asia. The most important advantages of smart cards are the capability and security that they offer.

One advantage of smart card IDs is they are extremely hard to forge. A PIN-code is added as an extra security

Figure 3. Dual slot mobile phone

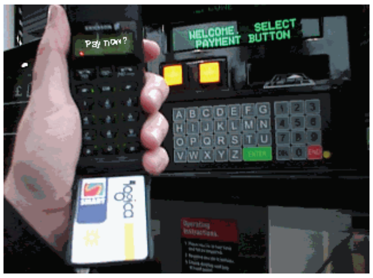
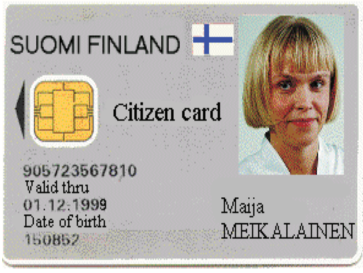


Figure 4. Smart card



measure to avoid abuse if the card gets stolen or lost. An ordinary ID card can only be used for identification, while a smart-card-based ID card can also be used to digitally sign documents and transactions in a non-repudiateable way (Rundgren, 1999a).

Palm Pad Computer and Laptop Computers

Palm pad and laptop computers have become another means of doing business over the Internet. Primarily developed to be portable and used for computing on the go, they are now being used for communication and access of information from databases at other locations. They originally could connect to the Internet via a mobile phone and conduct business. While laptops are fairly expensive, the palm pads are less expensive in price and are becoming more common in mobile commerce.

Software

There are four basic components that make up a wireless Web service: browser phones, WML, link server, and services. Browser phones are handheld devices with special software that replaces conventional Web browsers. The WML (Wireless Markup Language) is a programming language consisting of a set of statements that defines what the browser phone displays in its window and how it interacts with the user. Instead of Web pages, the wireless world uses decks consisting of cards (Vujosesevic & Laberge, 2000).

In 1995, European telecommunication companies wanted a common platform and decided on Java, which has today become the development language of choice for advanced cellular mobile phone services under the global system for mobile communication (GSM) digital communication platform. The advent of Java for the smart card computing environment, standardized as JavaCard API (application programming interface), now offers the prospect of an open mobile platform: one that can store multiple applications, as well as delete, replace, and upgrade them over the air, at the point-of-sale, or via the Internet. This technology gives operators new freedom to forge links with content providers, as well as develop their own unique applications and services (Brokat, 2000).

Wireless application protocol (WAP) is a leading global standard for delivering information over wireless devices. WAP bridges the gap between mobile devices and the Internet, delivering a wide range of mobile services to subscribers independent of their network, bearer, and terminal. The WAP-framework will be useful in digital cell phones in two ways: as a low-level communication protocol, and as an application environment supporting a "mini-browser." WAP is similar to the combination of HTML and HTTP, but includes optimization for low-bandwidth, low-memory, and low-display capability environments necessary to deliver information to mobile devices (Schone, 2004).

ISSUES AND CHALLENGES

Wireless Constraints

Developing content for wireless devices requires rethinking the Web experience. Wireless content developers need to begin from the ground up developing content for these new devices. These devices tend to have very little real estate available for viewing content—often as small as 14 × 7 characters. Wireless devices also tend to be monochromatic, so images do not render well. Keyboards are difficult to use. Wireless devices tend to have limited CPU, memory, and battery life. Developers and designers need to find new, intuitive navigational techniques to overcome these constraints. Today, the most common navigational technique on wireless is the drill-down capabilities (Gutzman, 2000).

Another constraint of wireless capabilities is the amount of bandwidth available for use of data transmission. This new technology would put a greater burden on current bandwidths available for wireless transmissions. Alternate bandwidths must be opened for transmission.

Wireless User Behavior

Wireless users will not be expected to "surf the Web" in the traditional sense. This is due to the viewing and input constraints of using a wireless device and the relative inconvenience of performing any but the most straightforward, time-critical tasks. More likely, wireless users are expected to use their devices to execute small, specific tasks that they can take care of quickly, such as finding the time of local events, purchasing tickets, looking up news, or checking e-mail. Content developers need to develop with these motives in mind. Rather than just translating a content-rich site into WML, developers need to think in terms of surgical access to content and drilling down capabilities to detailed information in the site (Gutzman, 2000).

Larger screens have been developed for viewing, but use of magnification or projection techniques would be easier for users to view Internet content. Keypads designed for smaller appliances should be developed with small typing ability in mind.

Infrastructure for Wireless Internet

Currently, the infrastructure to handle smart cards is not generally established (except in Europe). Most industry analysts believe that smart cards will eventually become mainstream for paying in shops and on the Internet together with a PC. In many countries, smart ID cards will also become fairly widespread. One of the problems is that the cost for shops, banks, companies, homes, and PC owners to convert to smart cards makes the process fairly slow. There is an obvious risk that consumers, banks, and companies, after the initial WAP euphoria is gone, start to question the rationale behind having multiple payment systems and could begin to put pressure on the mobile-phone makers to force them to adapt their systems to the rest of the world. This is a very awkward solution because it sets unnecessary physical constraints for mobile phones and is also likely to need "software fixes" for each new card variant. Even when used over GSM, operators will simply be supplying a gateway to the Internet, which will be regarded as a standard part of a subscription (Rundgren, 1999c).

Security

Security of data transmissions and commerce being conducted by wireless devices is a great concern for businesses and individuals today. The wired Internet is vulnerable to hacking attacks. Individuals have been wary of using Internet commerce for fear of having their credit card being used improperly. A prerequisite for the success of m-commerce applications is the legal recognition and non-disputability of any transactions affected. The mobile digital signature may be an answer to this problem (Brokat, 2000).

New smart cards, available for wireless communications applications, will enable secure transactions via the Internet. The wireless identity module (WIM) will guarantee a new level of security by giving mobile Internet users the ability to safeguard their transactions through encryption and digital signatures. Compliant with the wireless application protocol (WAP), the WIM device will allow mobile network operators and service and content providers to begin implementing mobile commerce services such as secure information access, online banking, and the purchase of goods and services. The WAP-powered identity module supports "logical channels," enabling users to pass from one application to another without losing transactions that have already been carried out. The card offers two forms of protection: client-to-server authentication using ultra-long keys, and the ability to generate the digital signature required to secure the application. Unlike an encryption-enabled browser, the secret keys handling the encryption remain in the user's smart card, by definition a tamper-resistant device, and allow it to be removed and transferred to other devices (Electronic, 1999).

One advantage of smart card IDs is they are extremely hard to forge. To crack a private key stored in such a smart card or guess its value based on corresponding public key is very difficult. A PIN-code is added as an extra security measure to avoid abuse if the card gets stolen or lost. An ordinary ID card can only be used for identification, while a smart-card-based ID card can also be used to digitally sign documents and transactions in a non-repudiateable way (Rundgren, 1999a).

Privacy

Privacy is another issue not resolved by the growth of mobile commerce. The new connectivity of consumers to the Internet is a great convenience for consumers, but it also comes at a price. The price is the value of privacy that individuals lose as they become hooked up to the Internet. One part of privacy is that the development of smart cards for use with cell phones is convenient for consumers wanting to buy or sell. However, much personal data is enclosed on the card, and it could be used for the wrong purposes. Many cell phones can be equipped with a global positioning chip, which can identify the location of the user. This new technology would be good for emergencies, but could also be used against the individual for monitoring purposes or other activity. These are issues that still need to be addressed and have been downplayed by current technology developers. Privacy is one of several issues that complicate the long-term timetable for developing location-based m-commerce. Another issue is the level of direct access marketers will have to customers since the Internet will be located with the individual customer and can be contacted by voice, e-mail, or Internet (Vujovesic & Laberge, 2000).

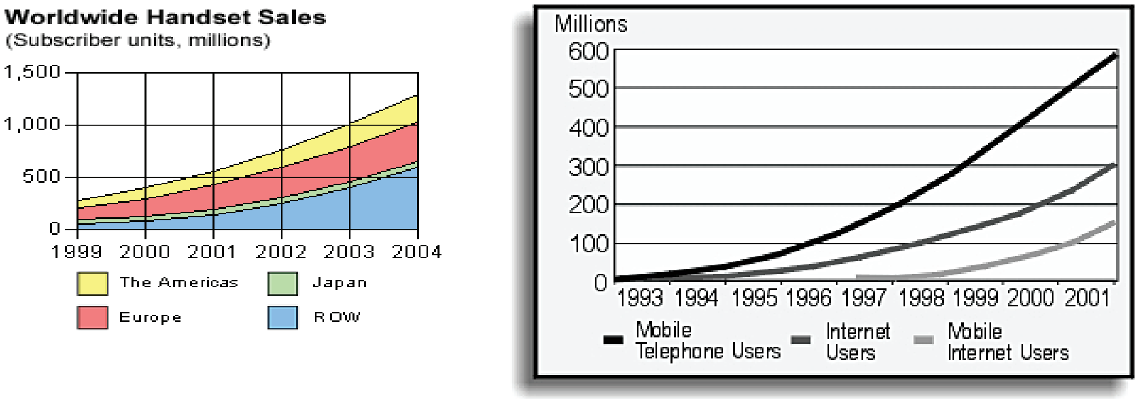
FUTURE TRENDS

The value of transactions conducted over mobile phones in Europe is set to reach 23 billion euros ($23.7 billion) by 2003, according to a new study from Durlacher Research. Europe's mobile phone service operators are poised to increasingly derive revenue from Internet content and services, and will become leading Internet portals in the future. Europe has adopted a clear lead in usage and application development, fueled by its high penetration of mobile phones and successful adoption of GSM as the single digital phone standard. The U.S. has not been able to reach a single standard nor to settle on a generic type of terminal, thus slowing the establishment of a critical mass of handsets in the market needed for introduction of new services (Uimonen, 1999).

According to a recent study by IDC, the number of people using wireless devices to connect to the Internet will increase by some 728% by 2003. That is an increase from 7.4 million users in 1999 to 61.5 million users in 2003 (Blackwell, 2000).

Forecasters predict that in 2003 over half of all Web access will be from a mobile device, by which time consumers will be comfortable with m-commerce. In 2003 around a quarter of all mobile Internet users are likely to use their mobile phone to access travel services such as booking flights,

Figure 5. Worldwide mobile commerce use (Cahners In-Stat Group)



finding local hotel accommodations, sourcing last-minute holidays, or purchasing rail tickets. These are all services that are particularly suited to both business and experienced leisure travelers. This mix of mobile transactions is likely to result in mobile travel commerce revenues overtaking online travel (Cross, 2000).

Smart cards have always offered the potential to radically change and automate the way mass consumer business operates. The huge business potential presented by Java-based smart cards to service providers has not been lost to GSM-based mobile communications operators. Today, the cost of mobile communications is almost the same as fixed-line communications. Mobile phone penetration in Asia is expected to reach 35-40%. With the merging of electronic tools, such as the palmtop with the mobile phone, and more and more using the mobile phone for accessing data, the mobile phone complements the Internet to enable Web surfers to access information without needing a PC. Eventually, e-commerce will be a hot application for mobile phone users. The same should be expected for GSM-based mobile e-commerce. The capability of mobile phones is expected to increase and this will accelerate more developments (Brokat, 2000).

IMPLICATIONS FOR MANAGEMENT

Consumer-oriented m-commerce is becoming a reality today. Many businesses and consumers are taking up the wireless Web through many services such as AT&T, Sprint PCS, Verizon Wireless, Motorola, Nokia, Ericsson, and other wireless service providers. It will still take a few years for consumer-oriented m-commerce applications to become as generally available as the wired Web is today. In the near term, the most promising opportunities for mobile wireless transactions are those built for industrial use. These involve the development of software for vertical applications that allow delivery agents, salespeople, and mobile workers to perform logistical and other data-driven duties. Doctors are using wireless-enabled palmtops to access and update patient records or write prescriptions. The most visible wireless developments are consumer oriented. These include delivery of time-critical information to mobile banking and travel-ticket purchase. The current crop of consumer-oriented services will become the foundation for more advanced services that will deliver time- and location-critical data to consumers (Vujovesic & Laberge, 2000).

Mobile commerce is a reality now and will not be going away in the near term. Problems with these systems are being addressed and new applications are being developed rapidly. Most of the market is behind this new technology, and it will likely change business by making it easier and more accessible to individuals. Mobile commerce is a tool of telecommunication and Internet industries. Those who are involved in it now may become the giants that Microsoft and Intel have been in the computing industry. Those who follow may be able to capitalize on leading m-commerce mistakes and perfect this technology. One should measure the risk involved and understand it must still be a carefully planned strategy to implement this new technology into corporate future goals.

CONCLUSION

The development of mobile commerce is the evolution of several different technologies to make the Internet more accessible and commerce easier for the consumer. While the Internet is already a valuable form of business, which has already changed the way the world is doing business, the format in which we will view it is changing. There are many new opportunities that have only begun to be explored. This will become a large opportunity for those who capitalize upon this technology. The growth trends are impressive, and the public interest and large companies are behind this technology.

Time will tell whether it is the treasure that most have touted it to be. If it is the "cash cow" most are looking for, then for those who trail or lag behind, the leaders may not be in business in a few years. If it turns out to be a bust, then those who invested so heavily in this technology will find their efforts wasted. The risk is very high in this new development, but the benefit is that consumers seem to be driving the demand for mobile commerce.

The application of this technology is the true seller. Its success is contingent upon a majority of Internet browsers using mobile digital phones. To be fully accepted, all these technologies must overcome their current drawbacks. Technology is being developed to overcome the security drawbacks, but enhanced viewing devices and input devices for controlling the data must be developed. Also, the infrastructure to control smart card payments may be a few years off for the U.S., but it will need to be accepted at shops and businesses throughout the U.S. to make it useful. Mobile e-commerce will change our world of business to a similar degree that the Internet alone has changed business today.

REFERENCES

Blackwell, G. (2000). *Wireless to outstrip wired Net access.* Retrieved from http:/www.isp-planet.com/research/more_wireless.html

Beal, A., Beck, J. C., Keating, S. T., Lynch, P. D., Tu, L., Wade, M., et al. (2001). *The future of wireless: Different than you think, bolder than you imagine.* Retrieved June 4, 2004, from http://www.accenture.com/xd/xd.asp?it=enWeb&xd=_isc/iscresearchreportabstract_134.xml

Cross, T. (2000). *Mobile travel commerce—A bigger deal than online travel?* Retrieved from http:/www.gmcforum.com/PressRelease/PressRelease_110500.htm

Durlacher Research. (1998). *Mobile electronic commerce.* Retrieved from http://network365.com/mobilecommerce.html

eTForecasts. (2006). *Internet user forecast by country: Wireless Internet users.* Retrieved on March 26, 2006, from http://www.etforecasts.com/products/ES_intusersv2.htm#1.0

Electronic Buyer's News. (1999). *Schlumberger says new smart card will ensure secure mobile-Internet transactions.* Retrieved from http:/www.ebns.com/ecomponents/commnews/story/OEG19991116S0008

Gartner Research. (2004, October 29). *Predictions 2005: Mobile and wireless technologies.* Retrieved July 3, 2005, from http://www.analysphere.com/13Aug01/wireless.htm

Gutzman, A. (2000). *The who, what and why of WAP.* Retrieved from http:/www.allnetdevices.com/wireless/opinions/2000/06/20/the_who.html

Hansen, C. (2000). *GSM-based mobile e-commerce will be hot.* Retrieved from http:/www.globalsources.com/MAGAZINE/TS/9909/SLB.HTM

Intel. (1999). *The future GSM data knowledge.* Retrieved from http:/www.gsmdata.com/Future.html

Magura, B. (2003). What hooks m-commerce customers? *MIT Sloan Management Review,* (Spring), 9.

Mobile Business. (2000). *Brokat global e-commerce services.* Retrieved from http:/www.brokat.com/int/mobile/index.html

Muller, J., & Schnoring, T. (1995). *Mobile telecommunications: Emerging European markets* (p. 247). Artech House Publishers.

Reiter, A. (1999a). *Dynamics of wireless e-commerce, conditions sparking the development of international wireless e-commerce.* Retrieved from http:/www.wirelessinternet.com/dynamics.htm

Reiter, A. (1999b). *Wireless e-commerce: A new business model.* Retrieved from http:/www.wirelessinternet.com/wireless2.htm

Rundgren, A. (1999a). *ID-cards: Yesterday, today and in the future.* Retrieved from http:/www.mobilephones-tng.com/papers/idcards.html

Rundgren, A. (1999b). *The cyber ID card.* Retrieved from http:/www.mobilephones-tng.com/v100/cyberphonecards.html

Rundgren, A. (1999c). *The new Swiss Army Knife? (Smart cards vs. smart terminals).* Retrieved from http./www.mobilephones-tng.com/papers/thenewswissarmyknife.htm

Rundgren, A. (1999d). *WAP—Wireless Application Protocol.* Retrieved from http:/www.mobilephones-tng.com/v100/wap.htm

Schone, S. (2004). *Computer Technology Review, 24*(10), 1, 38.

Strategis Group. (2005). *Mobile computing outlook.* Retrieved July 5, 2005, from http://www.mobileinfo.com/Market/market_outlook.htm

Uimonen, T. (1999). *European mobile commerce to hit $24 billion.* Retrieved from http:/www.durlacher.com

Varshney, U., & Vetter, R. (2000). Emerging mobile & wireless networks. *Communications of the ACM, 43*(6).

Vujosevic, S., & Laberge, R. (2000). *Info on the go: Wireless Internet database connectivity with ASP, XML, and SQL server.* Retrieved from http:/www.msdn.microsoft.com/msdnmag/issues/0600/wireless/wireless.asp

Walker, M. (2000). *M-commerce tricks emerge from tech magician's bag.* Retrieved from http:/www.bizjournals.com/houston/stories/2000/06/26/focus6.html

Wireless LAN Association. (2005). *Wireless LAN ROI.* Retrieved June 25, 2005, from http://wlana.org/learn/roi.htm

KEY TERMS

Mobile Commerce (M-Commerce): The delivery of electronic commerce capabilities directly into the consumer's hand via wireless technology and putting a retail outlet in the customers' hand anywhere.

Smart Card: A cross between an ID card and an electronic wallet. It can be used to store and exchange money from banks, as well as support the payment functions of digital cellular phones.

Wireless Application Protocol (WAP): A leading global standard for delivering information over wireless devices. It bridges the gap between mobile devices and the Internet, delivering a wide range of mobile services to subscribers independent of their network, bearer, and terminal.

Wireless Identity Module (WIM): Guarantees a new level of security by giving mobile Internet users the ability to safeguard their transactions through encryption and digital signatures.

Wireless Web Service: The four basic components that make up a wireless Web service are browser phones, WML, link server, and services.

Mobile Enterprise Readiness and Transformation

Rahul C. Basole
Georgia Institute of Technology, USA

William B. Rouse
Georgia Institute of Technology, USA

INTRODUCTION

Recent studies claim that mobile information and communication technologies (ICT) offer a plethora of new value propositions and promise to have a significant transformational impact on business processes, organizations, and supply chains (Kornak et al., 2004). However, despite its potential contributions, enterprise adoption of mobile ICT has not been as widespread as initially anticipated. Previous research has argued that successful adoption and implementation of any emerging ICT, such as mobile ICT, often requires fundamental changes across an enterprise and its current business practices, organizational culture, and workflows (Taylor & McAdam, 2004; Rouse, 2006). Hence, in order to minimize organizational risks and maximize the potential benefits of mobile ICT, enterprises must be cognizant of the value of enterprise mobility to their organization and accurately evaluate their level of "readiness" for mobile ICT adoption (Hartman & Sifonis, 2000; Ward & Peppard, 2002). This paper reviews the transformational value and impact of enterprise mobility and explores the critical dimensions for determining an enterprise's readiness for mobile ICT. Both theoretical and managerial implications are discussed.

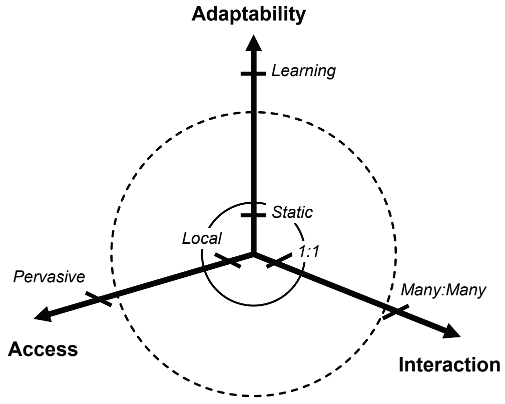
BACKGROUND

Over the past few years mobile ICT have advanced at a tremendous pace making an always-on connection, anywhere and anytime, a growing reality. The rapid proliferation of mobile devices has led to an increasingly mobile society in which users now expect to have instant communication means, data access, and commerce capabilities. A similar trend has also seeped into the enterprise domain. The use of mobile ICT in enterprises has evolved from being simplistic point solutions and small projects focused on productivity improvements and costs savings to strategic and large-scale enterprise-wide implementations that enable organizations to create new core competencies, gain and sustain competitive advantages, and define new markets (Davidson, 1999; Kornak et al., 2004).

The Mobile Enterprise

So what is a mobile enterprise? Simply deploying laptops so employees can take work home does not constitute a mobile enterprise. Pundits have argued that a slight increase in mobility that a laptop affords amounts to little more than a very small geographic extension of the existing static enterprise. Similarly, a mobile enterprise is not merely a collection of people with handheld devices, smart phones, tablet PCs, and pagers. Many enterprises already have such a workforce, however, it often does not change how those people work with each other and the rest of the organization. Therefore, bolting a group of mobile workers onto an organizational chart does not create a new organization and often does very little to enhance the existing one. However, the more mobile workers an organization has, the greater will be the need to transform at least part of that company into a mobile enterprise. More specifically, it will require a rethinking of how business is organized, how people interact and collaborate, how corporate resources are accessed, and how adaptable an enterprise is (Barnes, 2003; Rouse, 2005). Building on this notion, we propose that mobile enterprises exhibit higher levels of access, interaction, and adaptability than their static counterparts do. In visual terms, static enterprises tend to exist in spheres closer to the origin (see Figure 1). The further the sphere is from the origin, the higher the

Figure 1. The dimensions of the mobile enterprise



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

level of enterprise mobility. Thus, independent of location, the mobile enterprise is built on a foundation of processes and technologies allowing full access to organizational resources, which results in improved adaptability, access, and interaction among employees, customers, partners, and suppliers (Basole, 2005).

Benefits of Enterprise Mobility

With this understanding of the mobile enterprise, it therefore becomes more transparent what benefits mobile ICT can offer. The ability to access the corporate network and resources anywhere and anytime is one of the primary benefits and key drivers to adopting mobile enterprise solutions. Field workers are no longer tied to desktop computers to check mission- and task-critical data. The use of mobile ICT enables workers to receive timely answers, which in turn can lead to timely decisions. Enterprise mobility solutions also offer the potential of achieving significant cost savings. Expensive computing equipment can be replaced with smaller, more portable, and less expensive handheld devices. Field workers can use these devices to be immediately connected to all the sources they need. Furthermore, replacing paper-based processes with mobilized applications reduces the potential for errors in transferring information to a call report or clinical chart, leading to a higher level of data accuracy and integrity, which in turn can be harvested for overall business intelligence use. Better access to corporate resources—both data and people—naturally leads to a higher level of productivity, as mobile workers are able to view data that allows them to respond and execute faster to changing market conditions.

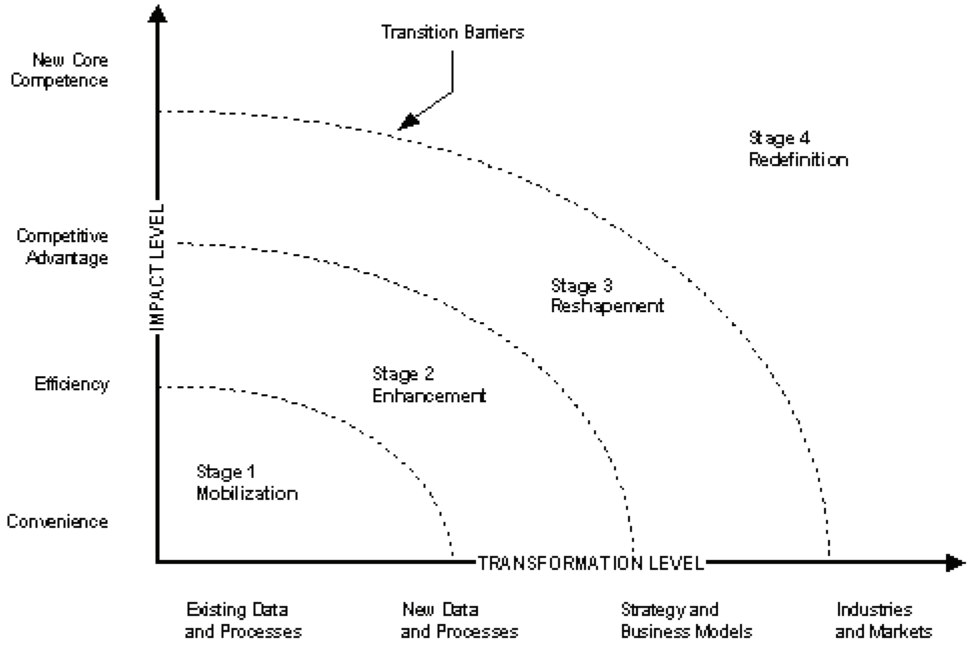
ENTERPRISE TRANSFORMATION THROUGH MOBILE ICT

Mobilizing enterprise applications and providing business professionals access to information anywhere and anytime is clearly an important first step in gaining business value (Barnes, 2003; Kornak et al., 2004); however these gains are only the beginning. We argue that enterprises can realize a much broader range of benefits over time by pursuing a multi-stage mobile transformation process. Research has shown that ICT have the ability to change and fundamentally transform enterprises in a number of ways (Basole & DeMillo, 2006; Rouse, 2006). This transformational impact can be primarily experienced and realized at the strategic, operational and organizational culture level (Taylor & McAdam, 2004). Indeed, the impact of mobile ICT is far beyond mere business process improvements and enhancements (Davidson, 1999; Kornak et al., 2004; Basole, 2005). Extending this previous work, four distinct stages of mobile enterprise transformations are proposed (see Figure 2).

Mobilization (Stage 1)

The first stage of the transformation process begins with the mobilization of existing data and applications. Mobilization refers to the process of making current business data, pro-

Figure 2. Stages of enterprise transformation through mobile ICT



cesses, and applications available for use on mobile/wireless devices. The first stage aims to provide end-users with a new level of convenience by enabling access to resources anywhere and anytime. Examples include access to corporate e-mail, the Intranet, and other data and human resources. Generally, Stage 1 solutions will lead to higher levels of convenience and generate significant performance gains in productivity, speed, efficiency, quality, and customer service (Kornak et al., 2004).

Enhancement (Stage 2)

The second stage shifts its focus from mobilizing existing data and applications to enhancing existing and creating new business processes that leverage the unique functionalities and capabilities of mobile ICT (Barnes, 2003). Characteristics of these business processes generally include two elements, namely (1) mobility (do it anywhere) and (2) immediacy (do it now), all with the user's context in mind. While solutions in the enhancement stage may affect working practices and modify business processes, they seldom change the business in a fundamental manner. This level of transformation occurs in Stage 3 of the mobile transformation process.

Reshapement (Stage 3)

As enterprises transition to Stage 3, mobile ICT begin to reshape business models and strategies. The creation of innovative new mobile processes and services provide enterprises with a source of competitive advantage. In this stage, mobile ICT often enable a business capability and become a critical element in the overall business model. For example, wireless sensors could enable a pharmaceutical company to shift from selling only medication to a business model in which the company provides both medication and sensors, and enters into a contract with a medical practitioner to perform continuous monitoring and keep a patient's blood pressure within an agreed range.

Redefinition (Stage 4)

In the fourth and final stage of the transformation process, mobile ICT create entirely new core enterprise competencies. Business models and strategies are based and revolve around enterprise mobility and in turn lead to a redefinition of entire markets and industries. Concrete examples for this stage of the mobile transformation process have not emerged yet; however, as enterprises continue to embrace mobility and mobile ICT mature, mobile redefinition is expected to become an increasingly common business phenomenon.

The four stages of mobile enterprise transformation are not purely sequential. Activities performed during Stage 1 continue during Stages 2-4. Some companies may elect transitioning directly from Stage 1 to Stage 3. New ventures may begin their business models based on Stage 2 philosophies. Stage 4 examples are still scarce, but are poised to emerge as mobile ICT continue to mature and new business models take shape. Yet, all four stages are inextricably linked in significant ways. Diligent pursuit of Stage 1 initiatives will lead to many Stage 2 and 3 opportunities. Similarly, Stage 4 opportunities will emerge as enterprises realize the full transformational potential of mobile ICT solutions.

Adoption and Transition Barriers

Enterprises that undergo significant organizational changes generally encounter a number of transition barriers. Empirical evidence suggests that these barriers can be broadly categorized as economic/strategic, technological, organizational, and environmental-related issues (Taylor & McAdam, 2004).

Despite tremendous advances, mobile ICT are still in their infancy stage. Evolving standards, lack of technology maturity, and issues of compatibility with existing systems and infrastructure are causing organizations to delay mobile ICT implementation (Basole, 2005). Another prevalent barrier is related to the ongoing debate of business value and cost. Investments in emerging technologies such as mobile ICT often require significant financial commitments by the enterprise. With shrinking IT budgets, it becomes critical to understand what value enterprise mobility can deliver now, and in the future. Mobile ICT implementations must thus be aligned with the overall business strategy and support enterprises' current and future business objectives. Similarly, the availability of other organizational resources – such as human and technical support—must be in place in order to successfully adopt mobile ICT and transition across the stages. From an organizational perspective, enterprise culture, size and structure also play a critical role in the adoption and transition process. As with most new ICT implementations, end-users often show resistance to new processes and change. Mobile ICT will have a radical, and potentially transformational, impact on the way work is done; hence, particular attention to end-user needs, education, motivation and incentives must be provided to ensure a successful adoption, implementation, and transition. Lastly, unfavorable market conditions, strong regulatory influences, lack of customer and supplier pressure, and inadequate vendor support may also inhibit organizational adoption of mobile ICT.

In summary, in order to avoid a "fragmented" mobile ICT adoption and transformation, enterprises should determine the fit between the value of mobile ICT and the overall business strategy, and ensure that a common vision, leadership support, and strategic path for implementing enterprise mobility solutions is in place (Ward & Peppard, 2002).

ENTERPRISE READINESS FOR MOBILE ICT

The previous discussion highlights the complexity of mobile ICT adoption and implementation, as well as the transition across the four transformation stages. It also underlines the fact that successful assimilation will require significant organizational changes, its current business practices, culture, processes, and workflows. While previous studies identified the potential benefits, challenges, and drivers for enterprise adoption of mobile ICT, exact enterprise benefits and ROI of mobile ICT implementations may not be known for the near future. As ICT budgets have decreased and failure rates for new ICT implementations have continued to rise over the past years, a smaller tolerance to ICT failures has emerged. The hype of potential benefits often drives enterprises to jump onto the "fad" bandwagon and rush into ineffective implementations of new ICT. In contrast, however, a range of studies have shown that many potentially successful IT projects fail due to a lack of assessment of potential barriers and organizational risks associated to the implementation of new ICT.

In order to minimize the associated risks and maximize the potential benefits of enterprise mobility solutions, organizations must thus not only understand the value and economics of enterprise mobility solutions, but also carefully evaluate and measure their level of "enterprise readiness" for mobile ICT (Hartman & Sifonis, 2000; Basole, 2005). Readiness assessment enables decision makers to become more knowledgeable about the characteristics of mobile ICT, form attitudes about it, and make a decision regarding the fit between the technology and the organization (Hartman & Sifonis 2000). It also enables decision makers to determine whether enterprises can truly benefit from mobile ICT and take appropriate measures to steer the organization towards a successful adoption and mobile transformation transition.

Defining Enterprise Readiness for Mobile ICT

The concept of enterprise readiness for ICT has received very limited attention in both the academic and business press literature. Preparedness, agility, and maturity are often some terms commonly associated with enterprise readiness. However, anecdotal evidence has shown that higher levels of enterprise readiness generally lead to lower levels of innovation risk and more successful implementation outcomes. A similar argument can be transposed to the context of mobile ICT: higher levels of mobile ICT readiness leads to lower organizational risks and implementations that are more successful.

So what constitutes enterprise readiness for mobile ICT? Extending previous theories of organizational readiness and technology adoption (Hartman & Sifonis, 2000; Taylor & McAdam, 2004), we postulate the following definition:

Enterprise readiness for mobile ICT is an assessment of an organization's (1) preparedness, (2) potential, and (3) willingness to adopt and implement mobile ICT.

More specifically, preparedness refers to an organization's ability to adopt, diffuse, and assimilate mobile ICT; potential refers to an organization's processes, employee, and strategy that could benefit from mobile ICT; and willingness reflects the attitudinal orientation of leadership and employee towards adopting mobile ICT.

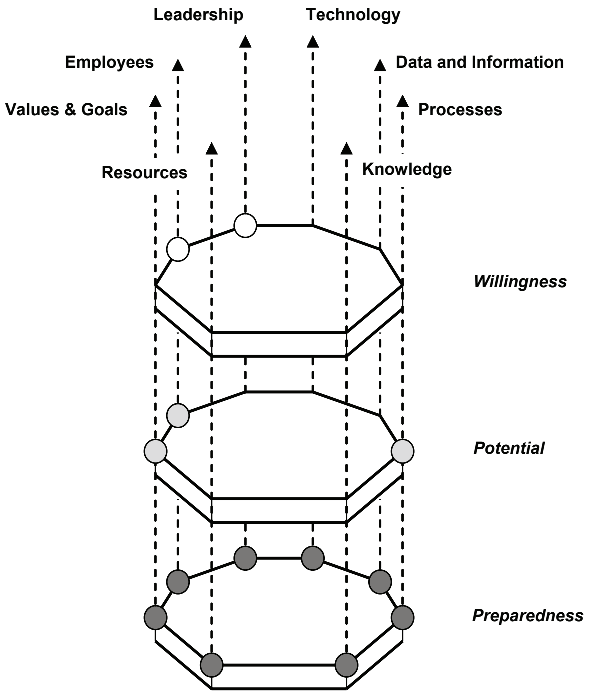
Dimensions of Enterprise Readiness

We further argue that enterprise readiness for mobile ICT is comprised of eight dimensions: (1) technology, (2) data and information, (3) process, (4) resource, (5) knowledge, (6) leadership, (7) employee, and (8) values and goals. A complete enterprise readiness assessment will thus involve an evaluation across the three layers—preparedness, potential, and willingness—and along all eight readiness dimensions (see *Figure 3*). Preparedness is assessed for all eight dimensions; potential is evaluated along the process, employee, and value and goals dimensions; and, willingness is assessed along the employee and leadership dimensions.

Theoretical and practical support for each of the eight dimensions and associated assessment indicators is provided as follows:

- **Technology Readiness:** Technology readiness refers to the ability of the underlying technology infrastructure (network services, hardware, software, and security) to support the adoption and implementation of mobile ICT. A robust, comprehensive, and open-standards oriented technological infrastructure, flexible and scalable to accommodate any change and emerging requirements, leads to a higher level of technology readiness.
- **Data and Information Readiness:** Data and information readiness refers to the ability to federate data from multiple sources, provide a single view of enterprise data, and make it available to any system at the time when it is needed. Higher levels of data and information readiness are achieved through a consistent, reliable, and secure data and information infrastructure that provides both synchronization and data recovery capabilities for highly disconnected and variable environments.
- **Process Readiness:** Process readiness refers to the ability of organizational processes (e.g., human processes, information processes, organizational change processes, etc.) to facilitate the adoption and imple-

Figure 3. Dimensions and layers of enterprise readiness for mobile ICT



mentation of mobile ICT. Well-defined, documented, managed, repeatable and optimized processes indicate a high level of readiness along this dimension.

- **Resource Readiness:** Resource readiness represents an organization's ability to support mobile ICT adoption and implementation. These resources may include (1) financial, (2) human, and (3) technical assets. The availability of resources for current and future plans is an important aspect in successful assimilations of mobile ICT.
- **Knowledge Readiness:** Knowledge readiness reflects both the general and specific knowledge required by decision makers for mobile ICT adoption and implementation. General knowledge includes awareness and understanding of the state of emerging ICT, ICT-related decision-making processes, and previous experiences with ICT adoptions and implementations. Specific knowledge encompasses an awareness and understanding of the opportunities, challenges, barriers, and opportunities that come with the adoption and implementation of mobile ICT. This will includes an understanding of mobile ICT characteristics, its potential impact on strategy, processes, and people, and the changing enterprise mobility market.
- **Leadership Readiness:** Previous studies have shown that one of the most critical factors in technology adoption decisions is the support and vision of top management. Leadership readiness, hence, reflects an appropriate level of skills, innovativeness, knowledge, and risk orientation of top management. It also indicates the level of support and strategic vision that management offers in association to the adoption and implementation of mobile ICT. Leadership needs to ensure that mobile strategies fit with the way they are doing business rather than changing their ways of doing business to fit the strategy.
- **Employee Readiness:** Employee readiness reflects the end-users attitude towards change, their level of skills, and perceived benefits by the end-users. A high level of employee readiness can lead to a faster adoption and diffusion of mobile ICT.
- **Values and Goals Readiness:** Values and goals readiness reflects the fit between existing structural and nonstructural enterprise characteristics and mobile ICT characteristics. Structural characteristics may include organizational size, centralization, formalization, autonomy, specialization, functional differentiation, strategic objectives and goals. Nonstructural characteristics may include culture, bureaucracy, task environment, and political climate.

It should be noted that all of these dimensions have an influence on each other and must therefore be considered as a whole. A lack in one dimension may influence the overall enterprise readiness for mobile ICT. Similarly, a lack of readiness in one of the three layers will also result in a lower degree of enterprise readiness. As such, a comprehensive assessment of all dimensions on all layers should be conducted.

FUTURE TRENDS AND CONCLUSION

Enterprise mobility is not merely a fad; it has become a reality in a wide-range of organizations and industries. Mobile ICT clearly offers a plethora of lucrative value propositions that will impact and fundamentally transform business processes, organizations, and supply chains. As mobile ICT continues to evolve and mature, enterprises must prepare themselves for a more "mobile" future. In order to minimize organizational risks and maximize the potential benefits of mobile ICT adoption and implementation, is therefore of utmost importance in order to assess the level of enterprise readiness for mobile ICT. We further argue that an understanding of the transformational stages can provide decision makers with a strategic map of the potential impact of mobile ICT. An assessment of an enterprise's preparedness, potential, and willingness in conjunction with an understanding of the transformative influence will enable decision makers to make more objective judgments on why and when to adopt mobile ICT, aid in the formulation of appropriate mobility strategies, and enable successful transformations.

REFERENCES

Barnes, S. J. (2003). Enterprise mobility: Concept and examples. *International Journal of Mobile Communications*, *1*(4), 341-359.

Basole, R.C. (2005). Mobilizing the enterprise: A conceptual model of transformational value and enterprise readiness. In *Proceedings of the 26th American Society of Engineering Management*. Virginia Beach, VA.

Basole, R. C., & DeMillo, R. A. (2006). Enterprise IT and transformation. In W. B. Rouse (Ed.), *Enterprise Transformation: Understanding and Enabling Fundamental Change* (pp. 223-237). New York: Wiley.

Davidson, W. (1999). Beyond re-engineering: The three phases of business transformation. *IBM Systems Journal*, *38*(2/3), 485.

Hartman, A., & Sifonis, J. (2000). *Net ready: Strategies for success in the e-conomy*, McGraw-Hill.

Kornak, A., et al. (2004). *Enterprise guide to gaining business value from mobile technologies*. New York: Wiley.

Rouse, W. B. (2005). A theory of enterprise transformation. *Systems Engineering*, *8*(4), 279-295.

Rouse, W. B. (2006). *Enterprise transformation: Understanding and enabling fundamental change*. New York: Wiley.

Taylor, J., & McAdam, R. (2004). Innovation adoption and implementation in organizations: A review and critique. *Journal of General Management*, *30*(1), 17-38.

Ward, J., & Peppard, J. (2002). *Strategic planning for information systems*. New York: Wiley.

KEY TERMS

Enterprise Readiness for Mobile ICT: Enterprise readiness for mobile ICT is an assessment of an organization's (1) preparedness, (2) potential, and (3) willingness to adopt and implement mobile ICT, and is an essential element for successful enterprise transformation through mobile ICT.

Enterprise Transformation: Enterprise transformation concerns fundamental change that substantially alters an organization's relationships with one or more key constituencies, and can involve new value propositions in terms of products and services, how these offerings are delivered and supported, and/or how the enterprise is organized to provide these offerings.

Mobile ICT: Mobile ICT refers to all mobile information and communication technologies, including network infrastructure (e.g., WiFi), devices (e.g., smart phones, laptops, PDAs), and mobile applications.

Mobile Entertainment

Chin Chin Wong
British Telecommunications (Asian Research Center), Malaysia

Pang Leang Hiew
British Telecommunications (Asian Research Center), Malaysia

INTRODUCTION

Mobile commerce is forecasted to be a significant growth market in leading countries. This high growth estimate of mobile phones is leading investors to take special interest in device manufacturing, provision for future innovations, and system management areas. Mobile commerce services can be adopted through different wireless and mobile networks, with the aid of several mobile devices (Andreou et al., 2002). Mobile commerce opens a new evolutionary era in global business (Maharramov, 1999). In mobile business there will be no need for international custom regulations that vary from country to country, therefore it is business without borders (Maharramov, 1999).

Mobile entertainment is a newly emerging subset of mobile commerce. A primary difficulty when researching mobile entertainment is that of definition (Moore & Rutter, 2004). It is recognized that, as mobile entertainment is a social and commercial process as well as a technical one, a diversity of other definitions for mobile entertainment is held by numerous industry producers, manufacturers, and end users, as well as researchers of dissimilar background (Moore & Rutter, 2004). It is noteworthy to rethink and redefine mobile entertainment, as it is more complex than other subsets of mobile commerce.

The problem of producing common understandings of mobile entertainment has been previously highlighted by the Mobile Entertainment Forum (MEF) when stating that two different industries make up the mobile entertainment industry: entertainment and telecommunications (Wiener, 2003). Mobile entertainment is created as the convergence of both industries. Each of these worlds speaks a different language and holds different assumptions about the nature of its work. Recent research demonstrates that many consumers are unclear about the mobile entertainment and related wireless technology options available to them. For example, a Packard Bell-sponsored survey of nearly 1,000 British home personal computer users found that 70% of the respondents did not know what Wi-Fi was (MORI, 2003).

Mobile entertainment represents one of the few mobile services that has mass market potential that will drive the adoption of the next generation of mobile devices (Ollila, Kronzell, Bakos, & Weisner, 2003). Proper classification of mobile entertainment services enable players in the value web to adopt suitable business models to bring services to market and how they should cooperate, share revenue, and jointly create competitive advantages.

This article presents a framework to examine mobile entertainment from multiple points of views concerning the service, network, and device-related sectors. This allows future research to be conducted with the clarity of distinguishing mobile entertainment services of different domains. The article also tries to collate and rationalize possibilities and restrictions of existing and emerging mobile entertainment technologies with respect to this framework. The study explores a number of scenarios to reflect the understanding on the value web. This study serves as a foundation for further studies in the area of mobile entertainment.

BACKGROUND

Travish and Smorodinsky (2002) as well as Kalyanaraman (2002) define mobile entertainment as services that offer gaming experiences on-par with those to be had in other mediums such as Xbox and PlayStation 2. On the contrary, it is of the authors' opinion that mobile entertainment services are more than merely games. Besides, the definition does not cover what constitutes mobile games. For example, if one considers games deployed on laptop and Game Boy as mobile games, a similar development approach could not be taken to launch mobile games on mobile phones because, generally, mobile games development on mobile devices should take into consideration key characteristics such as short session time, fresh content, continuous and reliable availability, cultural compliance, and so forth (Kalyanaraman, 2002). Furthermore, a game that is installed on a laptop cannot be installed on a mobile phone due to dissimilar platforms.

In other literature, Ollila et al. (2003) assume mobile entertainment includes any leisure activity undertaken via a personal technology, which is or has the potential to be networked, and facilitates transfer of data over geographic distance either on the move or at a variety of discrete locations. While workable, the definition does not cover whether mobile entertainment services must interact with service providers. It does not cover whether such service would

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Terminology of mobile entertainment (Wiener, 2003)

Terminology	Definition
Platform Vendor	Develops, implements, manufactures, supplies, and supports standard or customized platforms to the platform operator.
Service Provider	Brings content to the end user, undertakes the commercial and regulatory obligations that accompany the provision of service; does not involve the operator of the service.
Mobile Network Operator	Provides the infrastructure for mobile communications: the service, billing, and customer care.
Publisher	Refers to any company or individual that allows for the "publishing" of a piece of content; typically assumes the financial risk for the creation of the content; maintains control of all aspects of the entertainment service, including rights management and payment, user-service interaction, multi-user interaction, and user-per-service preferences.
Retailer	Delivers services to end users. In the mobile industry the retailers are either specialized for mobile services or mass retailers. Entertainment retailers are usually mass retailers.
Developer	Performs application development.
Subscriber	Refers to the end user or consumer of mobile entertainment services.

incur a cost upon usage. If mobile entertainment was said to be a subset of mobile commerce, it must therefore involve transaction of an economic value. The social aspects of mobile entertainment are hidden within the phrase "any leisure activity" (Moore & Rutter, 2004).

From a business perspective, various literatures attempt to classify the mobile entertainment value web by referring to its players within the industry. For example, Wiener (2003) asserts that to help all participants in this industry collaborate, clarification of how each industry defines the nature of its work is necessary. The goal is to offer a set of common definitions of typical industry players and various mobile entertainment roles for the interfaces between the businesses (Wiener, 2003). In another paper, Camponovo (2002) classifies the players in the value web based on technology, services, network, regulation, and user. A summary of the findings is concluded in Table 1.

A search on Google on the term *mobile entertainment* reveals that even everything portable, including DVD player, television, radio, external player, MP3 player, amplifiers, speakers, as well as woofers and so forth, are considered devices of mobile entertainment. This proves that confusion with regards to the definition of mobile entertainment is common among stakeholders of the value web.

Mobile entertainment comprises a range of activities including but not limited to downloading ring tone, logo, music, and movies; playing games; instant messaging; gambling; accessing location-based entertainment services; and Internet browsing. Hitherto, the list is constantly expanding.

REDEFINING MOBILE ENTERTAINMENT

In this section, the authors briefly explain the three different segments and come up with a model that is believed to be useful in the development of end user models and consumer scenarios. Subsequently, players in the mobile entertainment value web may improve their understanding of the consumers and their usage scenarios. This will make them perform better evaluations of the likelihood of adoption, and will improve their foundation for designing, evaluating, and timing mobile entertainment end user services (Pedersen, Methlie, & Thorbjørnsen, 2002).

In essence, taxonomy is a system of classifications. To put the framework into use, a few examples will be discussed in this section. The purpose of this section is to present a classification of these segments to identify relevant categories of mobile entertainment services for this study.

Scenario 1: Downloading Music onto Mobile Devices

A mobile user connects to the Internet via his 3G-enabled mobile phone, searches for a particular song, and downloads it onto his mobile phone. This falls under segment 1 where this activity utilizes wireless telecommunication networks, incurs a cost upon file download, interacts with the service provider, and is a form of leisure activity. If he transfers the music file to his friend via Bluetooth or infrared, this falls under segment 2 where such activity still utilizes the wireless network, yet does not incur a cost upon file transfer or involve any interaction with service providers. However, if he records his own singing (provided if the mobile device supports voice recording functionality), such activity is still considered as mobile entertainment, but it does not utilize the wireless network or incur a cost upon usage. Therefore, this activity falls under segment 3.

Scenario 2: Downloading Pictures onto Mobile Devices

A mobile user connects to the Internet via his 3G-enabled mobile phone, searches for a particular picture on the content provider's site, and downloads it onto his mobile phone as wallpaper. This activity falls under segment 1. If the mobile user transfers the image to his friend via Bluetooth or infrared, this activity falls under segment 2. On the other hand, if the mobile user snaps a picture with his camera phone, this would fall under segment 3. He did not download the picture, nor did he transfer the picture from another device.

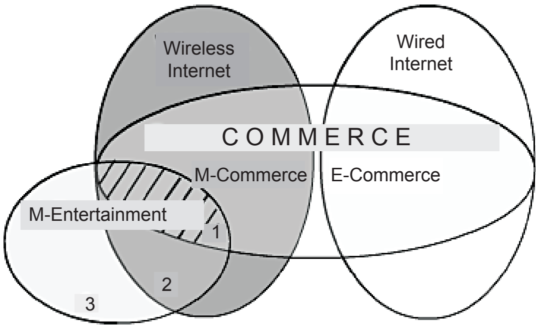
Scenario 3: Playing Games on Mobile Devices

A mobile user connects to the Internet via his WAP-enabled mobile phone, searches for a particular Java game, and downloads it onto his mobile phone. This activity falls under segment 1. Assuming that the downloaded game can be played as either a single player or multiplayer game, if the mobile user competes against his friend via Bluetooth, this activity falls under segment 2. In the third scenario, a mobile user plays preinstalled games on his mobile phone. He did not download the game nor did he transfer the game from another device. Therefore, the third scenario falls under segment 3.

Mobile entertainment does not exclude portfolio technologies such as Apple's iPod or Palm's Zire, which are not wireless but rely on being networked to other devices between periods of mobility (MGAIN, 2004). However, in the authors' opinion, such service falls under segment 3. Besides, such activity does not incur a cost and does not interact with service providers.

Hence, in these scenarios, the players in the value web vary in all three segments. The definitions of mobile entertainment for all three segments differ as well. Hence, the model in Figure 1 aids the industries to determine an appropriate business model to adopt in order to target the right audience. By classifying mobile entertainment service in its appropriate segment, it is then possible to determine the stakeholders involved, the network- and device-related requirements, and the business model required to develop and market the service (Wong & Hiew, 2005).

Figure 1. Mobile entertainment model: Position in the value web



FUTURE TRENDS

Mobile entertainment services have come a long way since the introduction of games like 'Snake' on mobile handsets (Drewitt & Bell, 2002). Consumers now have access to a host of services including online games, betting, and messaging, and the advent of 3G networks is poised to bring further developments (Drewitt & Bell, 2002).

Key players in the industry are taking the "wait-and-see" approach, as it is difficult to determine what the "killer applications" are in mobile entertainment arena. Some have already tested the waters with promotional music videos, sponsored ring tones, and song downloads (Williamson, 2005). But mobile video advertising could potentially be even bigger (Williamson, 2005).

However, there is a long list of challenges that will make mobile entertainment a difficult market to crack:

- **Bandwidth Issues:** Cellular systems like 2G and 3G have since been viewed as technologies capable of offering full mobility support due to their advanced roaming, handover, and network management capabilities, with relatively much lower data rate however. Yet, the so-called 3G mobile technology still lags in most of the United States (Williamson, 2005). Sprint and Verizon are the only two carriers to offer 3G phones (Williamson, 2005). Speed is the key driver for most mobile music and video applications.
- **Competing Technologies**: Mobile video today is an alphabet soup of technologies, and wireless operators have yet to agree on a standard (Williamson, 2005). That means content must be customized for each wireless network and consumers will, for the near term, select carriers based on the media content they offer (Williamson, 2005).

Finally, the mobile entertainment value web needs to carry on the innovation, development, and deployment of new technology that can be exploited for the improvement and expansion of value for mobile entertainment. This could be the means to finally make the consumers interested, as long as this is done from the consumers' perspective and not technology for its own sake. The future looks promising in this regard, with interesting new devices having additional functionality and improved user interfaces; and eventually, appealing 3G networks, security, and digital rights management (DRM) solutions will support a trustworthy use for both consumers and content providers.

CONCLUSION

The mobile revolution is changing the way people live and work. Mobile phones are already pervasive in all major developed economies and in an increasing number of developing ones as well. Prediction, based on both anecdotal and empirical information, on the future popularity and volume of mobile commerce has been widely presented in academic literature, as well as business and technological press.

In short, the authors define mobile entertainment as any leisure activity undertaken via a mobile device, interacting with service providers, incurring a cost upon usage, and utilizing wireless communication networks. Mobile entertainment services and applications can be adopted through different wireless and mobile networks, with the support of various mobile devices. An important factor in designing mobile entertainment services and applications is the necessity for apt identification of consumers' requirements, as well as mobile device and technology constraints. Services and applications are designed and developed based on these requirements and limitations.

Foresight is a series of methods and tools for creating future-orientated scenarios at national, regional, and sectoral levels. However, consumers are rarely consulted in traditional foresight exercises.

While consumer foresight is no more a game of forecasting the future than other forms of foresight research, it is clear that knowledge of consumer expectations can aid the mobile entertainment industry in focusing on those issues that are of concern to their market base. Directing technical and market developments towards fulfilling consumers' expectations of the future will support diffusion of new mobile technologies and services.

REFERENCES

Andreou, A. S., Chrysostomou, C., Leonidou, C., Mavromoustakos, S., Pitsillides, A., Samaras, G., et al. (2002). *Mobile commerce applications and services: A design and development approach, M-Business 2002.* Athens, July 8-9.

Booz, Allen, & Hamilton. (2003). Future mobile entertainment scenarios. *Proceedings of the Mobile Entertainment Forum,* (pp. 4-16).

Camponovo, G. (2002). *Mobile commerce business models, presented at International workshop on Business Models.* Lawsanne, Switzerland.

Drewitt, A., & Bell, P. (2002). *Play away: The future of mobile entertainment: BWCS.*

Kalyanaraman, R. (2002). *Mobile entertainment services—A perspective.* White Paper Series, Wipro Technologies.

Maharramov, S. (1999). M-commerce: Evolution in business. *Proceedings of the E-Business Forum,* pp. 1-6.

MGAIN. (2004). *Mobile entertainment industry and culture.* UK: MGAIN.

Moore, K., & Rutter, J. (2004). Understanding consumers' understanding of mobile entertainment. *Proceedings of Mobile Entertainment: User-Centred Perspectives,* Manchester, UK. pp. 113-148.

MORI. (2003). *Knowledge of WiFi hotspots.* MORI, London.

Ollila, M., Kronzell, M., Bakos, M., & Weisner, F. (2003). *Mobile entertainment industry and culture: Barriers and drivers.* UK: MGAIN.

Pedersen, P. E., Methlie, L. B., & Thorbjørnsen, H. (2002). Understanding mobile commerce end-user adoption: A triangulation perspective and suggestions for an exploratory service evaluation framework. *Proceedings of the 35th Hawaii International Conference on System Sciences,* Hawaii.

UK Trade and Investment. (2002). *Communication market in Malaysia, UK.*

Wiener, S. N. (2003). Terminology of mobile entertainment: An introduction. *Proceedings of the Mobile Entertainment Forum.*

Williamson, D. A. (2005). Mobile video: Present and future. *iMedia Connection,* (December 16). Available at http://www.imediaconnection.com/content/7587.asp

Wong, C. C., & Hiew, P. L. (2005). Mobile entertainment: Review and redefine. *Proceedings of the IEEE 4th International Conference on Mobile Business,* Sydney, Australia, pp. 187-192.

KEY TERMS

Bluetooth: A standard developed by a group of electronics manufacturers that allows any sort of electronic equipment to make its own connections, without wires or any direct action from a user. The name Bluetooth was derived from the tenth-century king of Denmark, known as King Harold Bluetooth, who engaged in diplomacy that led warning parties to negotiate with one another. The inventors of the Bluetooth technology thought this a fitting name for their technology, which allowed different devices to talk to each other.

Business Model: The mechanism by which a business intends to generate revenue and profits. It is a summary of how a company plans to select, serve, and keep its customers; define and differentiate its product offerings; position itself in the market; as well as capture profit. Also called a business design.

Consumer Scenario: Describes how consumers use a particular service or product in their everyday lives.

Digital Rights Management (DRM): The umbrella term referring to any of several technologies used to enforce predefined policies controlling access to software, music, movies, or other digital data. In more technical terms, DRM deals with the description, layering, analysis, valuation, trading, and monitoring of the rights held over a digital work.

Foresight: The providence by virtue of planning prudently for the future. In other words, it is the discipline of developing a forward view in time, the link-theme between spiritual dimension and mental dimension.

Infrared: Infrared data transmission is employed in short-range communication among computer peripherals and personal digital assistants. These devices usually conform to standards published by IrDA, the Infrared Data Association.

Mobile Commerce: A mobile commerce transaction is defined as any type of transaction of economic value that is conducted through a mobile device that uses a wireless telecommunications network for communication with the electronic commerce infrastructure.

Taxonomy: Initially taxonomy was only the science of classifying living organisms, but later the word was applied in a wider sense, and may also refer to either a classification of things or the principles underlying the classification.

3G: 3G is a short term for third-generation wireless, and refers to developments in personal and business wireless technology, especially mobile communications. 3G networks were conceived from the Universal Mobile Telecommunications Service (UMTS) concept for high-speed networks for enabling a variety of data-intensive applications.

Mobile File-Sharing over P2P Networks

Lu Yan
Åbo Akademi, Finland

INTRODUCTION

Peer-to-peer (P2P) computing is a networking and distributed computing paradigm which allows the sharing of computing resources and services by direct, symmetric interaction between computers. With the advance in mobile wireless communication technology and the increasing number of mobile users, peer-to-peer computing, in both academic research and industrial development, has recently begun to extend its scope to address problems relevant to mobile devices and wireless networks.

The mobile ad hoc network (MANET) and P2P systems share key characteristics including self-organization and decentralization, and both need to solve the same fundamental problem: connectivity. Although it seems natural and attractive to deploy P2P systems over MANET due to this common nature, the special characteristics of mobile environments and the diversity in wireless networks bring new challenges for research in P2P computing.

Currently, most P2P systems work on wired Internet, which depends on application layer connections among peers, forming an application layer overlay network. In MANET, overlay is also formed dynamically via connections among peers, but without requiring any wired infrastructure. So the major differences between P2P and MANET that concern us in this article are:

a. P2P is generally referred to the application layer, but MANET is generally referred to the network layer, which is a lower layer concerning network access issues. Thus, the immediate result of this layer partition reflects the difference of the packet transmission methods between P2P and MANET: the P2P overlay is a unicast network with virtual broadcast consisting of numerous single unicast packets, while the MANET overlay always performs physical broadcasting.
b. Peers in P2P overlay are usually referred to static node though no priori knowledge of arriving and departing is assumed, but peers in MANET are usually referred to mobile node since connections are usually constrained by physical factors like limited battery energy, bandwidth, computing power, and so forth.

BACKGROUND

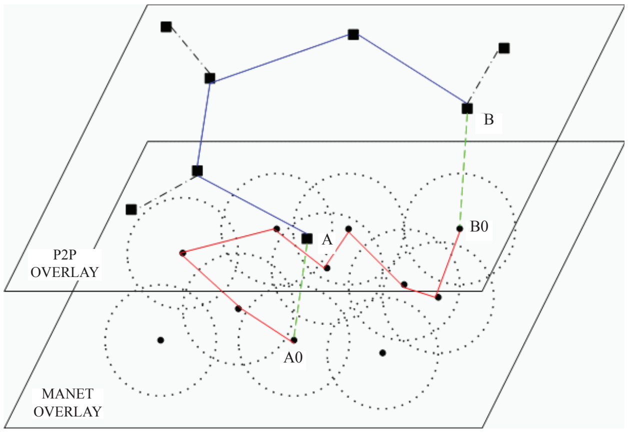
Since both P2P and MANET are becoming popular only in recent years, the research on P2P systems over MANET is still in its early stage. The first documented system is Proem (Kortuem et al., 2001), which is a P2P platform for developing mobile P2P applications, but it seems to be a rough one, and only IEEE 802.11b in ad hoc mode is supported. 7DS (Papadopouli & Schulzrinne, 2001) is another primitive attempt to enable P2P resource sharing and information dissemination in mobile environments, but it is rather a P2P architecture proposal than a practical application. In a recent paper, Lindemann and Waldhorst (2002) proposed *passive distributed indexing* for such kinds of systems to improve the search efficiency of P2P systems over MANET, and in ORION (Klemm, Lindemann & Waldhorst, 2003), a broadcast-over-broadcast routing protocol was proposed. The above works were focused on either P2P architecture or routing schema design, but how efficient the approach is and what the performance experienced by users is—these are still in need of further investigation.

Previous work on performance study of P2P over MANET was mostly based on simulative approach, and no concrete analytical mode was introduced. Performance issues of these kinds of systems were first discussed in Goel, Singh, and Xu (2002), but it simply shows the experiment results and no further analysis was presented. There is a survey of such kinds of systems in Ding and Bhargava (2004), but no further conclusions were derived. A sophisticated experiment and discussion on P2P communication in MANET can be found in Hsieh and Sivakumar (2004). However, all above works fall into a practical experience report category, and no performance models are proposed.

There have been many routing protocols in P2P networks and MANET respectively. For instance, one can find a very substantial P2P routing scheme survey from HP Labs in Milojicic et al. (2002), and U.S. Navy Research publishes ongoing MANET routing schemes (MANET, n.d.); but all of the above schemes fall into two basic categories: broadcast-like and DHT-like. More specifically, most early P2P search algorithms, such as in Gnutella (www.gnutella.com), Freenet (freenet.sourceforge.net), and Kazaa (www.kazaa.com), are

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Broadcast over broadcast



broadcast-like, and some recent P2P searching, like in eMule (www.emule-project.net) and BitTorrent (http://bittorrent.com/), employs more or less some feathers of DHT. On the MANET side, most on-demand routing protocols such as DSR (n.d.) and AODV (n.d.) are basically broadcast-like. Therefore, we here introduce different approaches to integrate these protocols in different ways according to categories.

BROADCAST OVER BROADCAST

The most straightforward approach is to employ a broadcast-like P2P routing protocol at the application layer over a broadcast-like MANET routing protocol at the network layer. Intuitively, in the above settings, every routing message broadcasting to the virtual neighbors at the application layer will result in a full broadcast to the corresponding physical neighbors at the network layer.

The scheme is illustrated in Figure 1 with a searching example: peer A in the P2P overlay is trying to search for a particular piece of information, which is actually available in peer B. Due to broadcast mechanism, the search request is transmitted to its neighbors, and recursively to all the members in the network, until a match is found or it times-out. Here we use the blue lines to represent the routing path at this application layer. Then we map this searching process into the MANET overlay, where node A0 is the corresponding mobile node to the peer A in the P2P overlay, and B0 is related to B in the same way. Since the MANET overlay also employs a broadcast-like routing protocol, the request from node A0 is flooded (broadcast) to directly connected neighbors, which in turn flood their neighbors and so on, until the request is answered or a maximum number of flooding steps occur. The route establishing lines in that network layer are highlighted in red, where we can find that there are few overlapping routes between these two layers, though they all employ broadcast-like protocols.

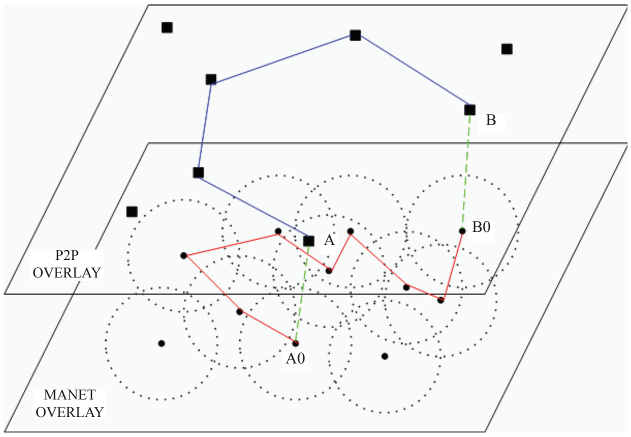
We have studied a typical broadcast-like P2P protocol, Gnutella (Clip2, 2001), in previous work (Yan & Sere, 2003). This is a pure P2P protocol, in which no advertisement of shared resources (e.g., directory or index server) occurs. Instead, each request from a peer is broadcast to directly connected peers, which themselves broadcast this request to their directly connected peers and so on, until the request is answered or a maximum number of broadcast steps occur. It is easy to see that this protocol requires a lot of network bandwidth, and it does not prove to be very scalable. The complexity of this routing algorithm is O(n) (Ripeanu, Foster, & Iamnitch, 2002; Chawathe, Ratnasamy, Breslau, & Shenker, 2003).

Generally, most on-demand MANET protocols, like DSR (Johnson & Maltz, 1996) and AODV (Perkins & Royer, 2000), are broadcast-like in nature (Kojima, Harada, & Fujise, 2001). Previously, one typical broadcast-like MANET protocol, AODV, was studied (Yan & Ni, 2004). In that protocol, each node maintains a routing table only for active destinations: when a node needs a route to a destinations, a path discovery procedure is started based on a RREQ (route request) packet; the packet will not collect a complete path (with all IDs of involved nodes) but only a hop count; when the packet reaches a node that has the destination in its routing table, or the destination itself, a RREP (route reply) packet is sent back to the source (through the path that has been set up by the RREQ packet), which will insert the destination in its routing table and will associate the neighbor from which the RREP was received as the preferred neighbor to that destination. Simply speaking, when a source node wants to send a packet to a destination, if it does not know a valid route, it initiates a route discovery process by flooding the RREQ packet through the network. AODV is a pure on-demand protocol, as only nodes along a path maintain routing information and exchange routing tables. The complexity of that routing algorithm is O(n) (Royer & Toh, 1999).

This approach is probably the easiest one to implement, but the drawback is also obvious: the routing path of the requesting message is not the shortest path between source and destination (e.g., the red line in Figure 1), because the virtual neighbors in the P2P overlay are not necessarily also the physical neighbors in the MANET overlay, and actually these nodes might be physically far away from each other. Therefore, the resulting routing algorithm complexity of this broadcast-over-broadcast scheme is unfortunately $O(n^2)$, though each layer's routing algorithm complexity is O(n) respectively.

It is not practical to deploy such a scheme for its serious scalability problem due to the double broadcast; and taking the energy consumption portion into consideration, which is somehow critical to mobile devices, the double broadcast will

Figure 2. DHT over broadcast



also cost a lot of energy consumption and make it infeasible in cellular wireless data networks.

DHT OVER BROADCAST

The scalability problem of broadcast-like protocols has long been observed, and many revisions and improvement schemas are proposed (Lv, Ratnasamy, & Shenker, 2002; Yang & Garcia-Molina, 2002; Chawathe et al., 2003). To overcome the scaling problems in broadcast-like protocols where data placement and overlay network construction are essentially random, a number of proposals are focused on structured overlay designs. The distributed hash table (DHT) (Stoica, Morris, Karger, Kaashoek, & Balakrishnan, 2001) and its varieties (Ratnasamy, Francis, Handley, Karp, & Schenker 2001; Rowstron & Druschel, 2001; Zhao et al., 2004) advocated by Microsoft Research seem to be promising routing algorithms for overlay networks. Therefore it is interesting to see the second approach: to employ a DHT-like P2P routing protocol at the application layer over a broadcast-like MANET routing protocol at the network layer.

The scheme is illustrated in Figure 2 with the same searching example. Compared to the previous approach, the difference lies in the P2P overlay: in a DHT-like protocol, files are associated to keys (e.g., produced by hashing the file name); each node in the system handles a portion of the hash space and is responsible for storing a certain range of keys. After a lookup for a certain key, the system returns the identity (e.g., the IP address) of the node storing the object with that key. The DHT functionality allows nodes to put and get files based on their key, and each node handles a portion of the hash space and is responsible for a certain key range. Therefore, routing is location-deterministic distributed lookup (e.g., the blue line in Figure 2).

DHT was first proposed in Plaxton, Rajaraman, and Richa (1997) without intention to address P2P routing problems. DHT soon proved to be a useful substrate for large distributed systems, and a number of projects are proposed to build Internet-scale facilities layered above DHTs; among them are Chord, CAN, Pastry, and Tapestry. All take a key as input and route a message to the node responsible for that key. Nodes have identifiers, taken from the same space as the keys. Each node maintains a routing table consisting of a small subset of nodes in the system. When a node receives a query for a key for which it is not responsible, the node routes the query to the hashed neighbor node towards resolving the query. In such a design, for a system with n nodes, each node has O(log n) neighbors, and the complexity of the DHT-like routing algorithm is O(log n) (Ratnasamy, Shenker, & Stoica, 2002).

Additional work is required to implement this approach, partly because DHT requires a periodical maintenance (i.e., it is just like an Internet-scale hash table or a large distributed database); since each node maintains a routing table (i.e., hashed keys) to its neighbors according to DHT algorithm, following a node join or leave, there is always a nearest key reassignment between nodes.

This DHT-over-broadcast approach is obviously better than the previous one, but it still does not solve the shortest path problem as in the broadcast-over-broadcast scheme. Though the P2P overlay algorithm complexity is optimized to O(log n), the mapped message routing in the MANET overlay is still in the broadcast fashion with complexity O(n); the resulting algorithm complexity of this approach is as high as O(n log n).

This approach still requires a lot of network bandwidth and hence does not prove to be very scalable, but it is efficient in limited communities such as a company network.

CROSS-LAYER BROADCAST

A further step of the broadcast-over-broadcast approach would be a cross-layer broadcast. Due to similarity of broadcast-like P2P and MANET protocols, the second broadcast could be skipped if the peers in the P2P overlay would be mapped directly into the MANET overlay, and the result of this approach would be the merge of application layer and network layer (i.e., the virtual neighbors in P2P overlay overlaps the physical neighbors in MANET overlay).

The scheme is illustrated in Figure 3, where the advantage of this cross-layer approach is obvious: the routing path of the requesting message is the shortest path between source and destination (e.g., the blue and red lines in Figure 3), because the virtual neighbors in the P2P overlay are de facto physical neighbors in the MANET overlay due to the merge of two layers. Thanks to the nature of broadcast, the algorithm complexity of this approach is O(n), making it

Figure 3. Cross-layer broadcast

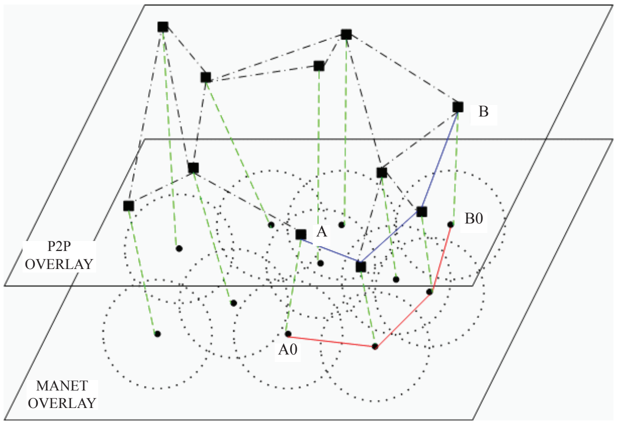
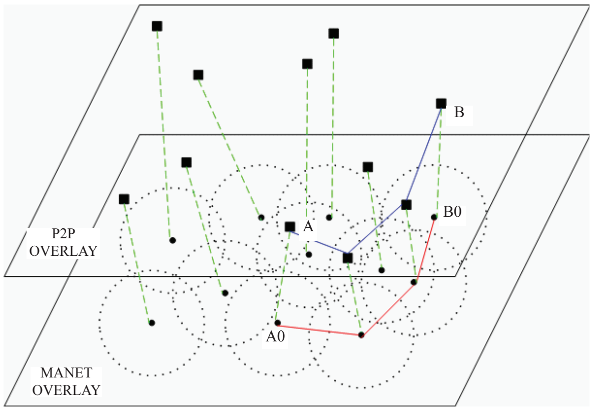


Figure 4. Cross-layer DHT



suitable for deployment in relatively large-scale networks, but still not feasible for Internet-scale networks.

CROSS-LAYER DHT

It is also possible to design a cross-layer DHT in Figure 4 with the similar inspiration, and the algorithm complexity would be optimized to O(log n) with the merit of DHT, which is advocated to be efficient even in Internet-scale networks. The difficulty in that approach is implementation: there is no off-the-shelf DHT-like MANET protocol as far as we know, though recently, some research projects like Ekta (Pucha, Das, & Hu, 2004) towards a DHT substrate in MANET are proposed.

CONCLUSION

In this article, we studied the peer-to-peer systems over mobile ad hoc networks with a comparison of different settings for the peer-to-peer overlay and underlying mobile ad hoc network. We show that the cross-layer approach performs better than separating the overlay from the access networks in Table 1. Our results would potentially provide useful guidelines for mobile operators, value-added service providers, and application developers to design and dimension mobile peer-to-peer systems.

REFERENCES

AODV. (n.d.). *AODV IETF draft v1.3*. Retrieved from http://www.ietf.org/internet-drafts/draft-ietf-manet-aodv-13.txt

Chawathe, Y., Ratnasamy, S., Breslau, L., & Shenker, S. (2003). Making Gnutella-like P2P systems scalable. *Proceedings of ACM SIGCOMM.*

Clip2. (2001). *The Gnutella protocol specification v0.4* (document revision 1.2). Retrieved from http://www9.limewire.com/developer /gnutella protocol 0.4.pdf

DSR. (n.d.). *DSR IETF draft v1.0*. Retrieved from http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt

Ding, G., & Bhargava, B. (2004). Peer-to-peer file-sharing over mobile ad hoc networks. *Proceedings of the* 2^{nd} *IEEE Conference on Pervasive Computing and Communications Workshops.*

Goel, S. K., Singh, M., & Xu, D. (2002). Efficient peer-to-peer data dissemination in mobile ad-hoc networks. *Proceedings of the International Conference on Parallel Processing.*

Table 1. How efficient does a user try to find a specific piece of data?

	Efficiency	Scalability	Implementation
Broadcast over Broadcast	O(n^2)	n/a	Easy
DHT over Broadcast	O(n log n)	Bad	Medium
Cross-Layer Broadcast	O(n)	Medium	Difficult
Cross-Layer DHT	O(log n)	Good	n/a

Hsieh, H.Y., & Sivakumar, R. (2004). On using peer-to-peer communication in cellular wireless data networks. *IEEE Transactions on Mobile Computing, 3*(1).

Johnson, D. B., & Maltz, D. A. (1996). *Dynamic source routing in ad-hoc wireless networks. Mobile computing.* Kluwer.

Klemm, A., Lindemann, C., & Waldhorst, O. (2003). A special-purpose peer-to-peer file sharing system for mobile ad hoc networks. *Proceedings of the IEEE Vehicular Technology Conference.*

Kojima, F., Harada, H., & Fujise, M. (2001). A study on effective packet routing scheme for mobile communication network. *Proceedings of the 4th Symposium on Wireless Personal Multimedia Communications.*

Kortuem, G., Schneider, J., Preuitt, D., Thompson, T. G. C., Fickas, S., & Segall, Z. (2001). When peer-to-peer comes face-to-face: Collaborative peer-to-peer computing in mobile ad hoc networks. *Proceedings of the 1st International Conference on Peer-to-Peer Computing.*

Lindemann, C., & Waldhorst, O. (2002). A distributed search service for peer-to-peer file sharing in mobile applications. *Proceedings of the 2nd IEEE Conference on Peer-to-Peer Computing.*

Lv, Q., Ratnasamy, S., & Shenker, S. (2002). Can heterogeneity make Gnutella scalable? *Proceedings of the 1st International Workshop on Peer-to-Peer Systems.*

MANET. (n.d.). *MANET implementation survey.* Retrieved from http://protean.itd.nrl.navy.mil/manet/survey/survey.html

Milojicic, D. S., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., Rollins, S., & Xu, Z. (2002). *Peer-to-peer computing.* Technical Report HPL-2002-57, HP Labs.

Papadopouli, M., & Schulzrinne, H. (2001). A performance analysis of 7DS, a peer-to-peer data dissemination and prefetching tool for mobile users. *IEEE Sarnoff Symposium Digest.*

Perkins, C. E., & Royer, E. M. (2000). *The ad hoc on-demand distance vector protocol. Ad hoc networking.* Boston: Addison-Wesley.

Plaxton, C., Rajaraman, R., & Richa, A. (1997). Accessing nearby copies of replicated objects in a distributed environment. *Proceedings of ACM SPAA.*

Pucha, H., Das, S. M., & Hu, Y. C. (2004). Ekta: An efficient DHT substrate for distributed applications in mobile ad hoc networks. *Proceedings of the 6th IEEE Workshop on Mobile Computing Systems and Applications.*

Ratnasamy, S., Francis, P., Handley, M., Karp, R., & Schenker, S. (2001). A scalable content-addressable network. *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications.*

Ratnasamy, S., Shenker, S., & Stoica, I. (2002). Routing algorithms for DHTs: Some open questions. *Proceedings of the 1st International Workshop on Peer-to-Peer Systems.*

Ripeanu, M., Foster, I., & Iamnitch, A. (2002). Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing, 6*(1).

Rowstron, A., & Druschel, P. (2001). Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms.*

Royer, E. M., & Toh, C. K. (1999). *A review of current routing protocols for ad-hoc mobile wireless networks.* IEEE Personal Communications.

Stoica, I., Morris, R., Karger, D., Kaashoek, F., & Balakrishnan, H. (2001). Chord: A scalable peer-to-peer lookup service for Internet applications. *Proceedings of ACM SIGCOMM.*

Yan, L., & Ni, J. (2004). Building a formal framework for mobile ad hoc computing. *Proceedings of the International Conference on Computational Science.*

Yan, L., & Sere, K. (2003). Stepwise development of peer-to-peer systems. *Proceedings of the 6th International Workshop in Formal Methods.*

Yang, B., & Garcia-Molina, H. (2002). Improving search in peer-to-peer networks. *Proceedings of the International Conference on Distributed Systems.*

Zhao, B. Y., Huang, L., Stribling, J., Rhea, S. C., Joseph, A. D., & Kubiatowicz, J. (2004). Tapestry: A Resilient global-scale overlay for service deployment. *IEEE Journal on Selected Areas in Communications.*

KEY TERMS

AODV: Ad hoc On-demand Distance Vector routing.

DHT: Distributed hash table.

DSR: Dynamic source routing.

MANET: Mobile ad hoc network.

P2P: Peer-to-peer.

Mobile Gaming

Krassie Petrova
Auckland University of Technology, New Zealand

INTRODUCTION

A number of multifunctional handheld devices with Internet and multimedia capabilities are currently available on the market. The mobile network technologies implemented make it possible for a range of value-added mobile services known as "mobile entertainment" to be offered to paying subscribers (Carlsson, Hyvonen, Repo, & Walden, 2005). Examples include watching streamed news, downloading music and images, or playing a game on one's mobile phone (alone, or in an interaction with other players).

Mobile game development depends on the choice of a middleware platform, as the application needs to be portable across a wide spectrum of handheld devices and technologies (Yuan, 2004; Hagleitner & Mueck, 2002). A business model for offering mobile gaming as a service has been successfully trialled in Japan where playing games is one of the main components of the popular Japanese entertainment platform iMode (Natsuno, 2003, pp.88-90).

Research in the area of mobile gaming adoption has focused on the investigation of the value generation process and on identifying the critical factors for mobile gaming acceptance. A number of critical success factors have been identified (e.g., Shchglick, Barnes, Scornavacca, & Tate, 2004; Moore & Rutter, 2004; Yoon, Ha, & Choi, 2005), adapting and extending existing mobile business frameworks and models (e.g., Siau, Lim, & Shen, 2001; Varshney & Vetter, 2002; Lee, Hu, & Yeh, 2003; Barnes, 2003). This short article investigates the relationship between mobile gaming customers and the mobile gaming value chain, and discusses the implications from mobile gaming supply and demand perspectives.

BACKGROUND

Playing mobile games ("mobile gaming") is classified as a "mobile entertainment" application (Barnes & Huff, 2003; Van de Kar, Maitland, de Montalvo, & Bouwman, 2003). Mobile entertainment includes personal leisure activities undertaken via a network technology. Entertainment services might feature data transfer including voice and video over significant geographic distance while the user of the service is either on the move or has the potential to move without interrupting the activity (Ollila, Kronzell, Bakos, & Weisner, 2003).

Mobile gaming is also an example of a mobile commerce (m-commerce) application, provided through a specially designed m-commerce service. Typically, an m-commerce service involves payment: a monetary transaction which the customer conducts using the mobile payment mechanism provided with the service (Paavilainen, 2002). In the case of mobile gaming, the player's subscriber account with the mobile network operating the service is used to collect the revenue. Subsequently the network operator makes payments to other service providers who might be involved: content developers and publishers, portal aggregators, and retailers (Wiener, 2003).

Some mobile games are simply downloaded and played off-line, paying once or with every update. Such games might be suitable for "low-end" handheld devices and might use text messaging. Other mobile games need to be played on smartphones, as interactivity among multiple players needs to be supported. These real-time mobile games require a persistent network connection to a dedicated game server. Advancements in mobile game development include the inclusion of location-based features into the game (Moore & Rutter, 2004; Maitland, van de Kar, de Montalvo, & Bouwman, 2005). In all cases, the mobile game player does not need to be stationary – he or she is "released" from the need to use a stationary networked device (Finn, 2005) .The assumed mobility of the mobile game player is one of the defining features of mobile gaming.

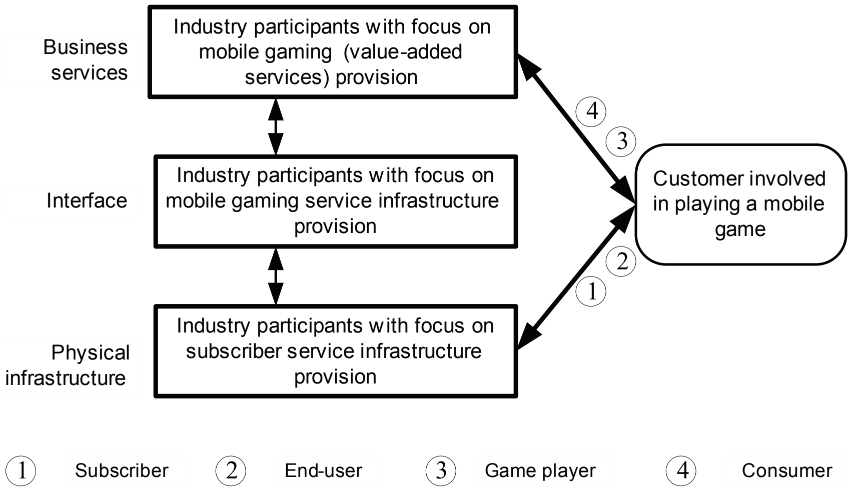
A leader in mass mobile entertainment, the Japanese company NTT DoCoMo developed a comprehensive mobile service and platform: iMode. Entertainment applications and specifically mobile gaming are seen as the catalyst for the increased use of the range of other iMode services (Baldi & Thaung, 2002; Natsuno, 2003, p.92; Barnes & Huff, 2003; Funk, 2003, p. 28). In the global market, mobile gaming is seen as a viable business opportunity (Kleijnen, de Ruyter, & Wetzels, 2003, p. 205; Anckar & D'Incau, 2002; Paavilainen, 2002). According to some predictions, by 2010 the revenue from downloading mobile games might reach US$8.4 billion (Graft, 2006).

THE MOBILE GAMING VALUE CHAIN

A number of value chain models (e.g., Buellingen & Woerter, 2004; Siau et al., 2001, Barnes, 2003) and mobile frameworks (e.g., Varshney & Vetter, 2002; Stanoevska-Slabeva, 2003)

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Customer relationships in mobile gaming (Derived from Petrova, 2005)


Business services
Industry participants with focus on mobile gaming (value-added services) provision
Interface
Industry participants with focus on mobile gaming service infrastructure provision
Physical infrastructure
Industry participants with focus on subscriber service infrastructure provision
Customer involved in playing a mobile game
④
③
②
①
① Subscriber
② End-user
③ Game player
④ Consumer

for m-commerce have been suggested in the literature and used to map industry players, roles and functions. A multiple value-chain model representing the relationships between a customer and the mobile gaming industry is shown in Figure 1. It applies the m-commerce reference model proposed by Petrova (2005) to the mobile entertainment value Web (Ollila et al., 2003) to identify the relationships between the customer (a mobile game player) and the mobile industry participants.

The network developers and providers, along with device developers, create the physical foundation needed for mobile gaming, and together with the mobile network providers, form a physical infrastructure layer. The interface layer includes developers of middleware platforms, which serve as game developing and servicing environments and enable the use of the networks and technologies for game service provision. The top layer represents the category of value-added services where mobile games are offered to customers directly or as a part of a mobile entertainment package. The main industry participants are mobile game developers, publishers and aggregators.

A company involved in mobile gaming can be categorised under more than one category: *Vodafone,* for example, provides a subscriber network and subscriber service, as well as mobile game downloading through its portal Vodafone Live! (Harmer, 2003). The revenue model of the company depends on its position in the value chain (Ollila et al., 2003).

MOBILE VALUE CHAIN RELATIONSHIPS

Mobile network operators are part of the physical infrastructure by building and maintaining the network. In most cases they also provide subscriber services and access to the network (e.g., *Vodafone, Orange).* The 2G-2.5G technologies currently implemented include CDMA, GSM, and GPRS, and are capable of maintaining real-time, persistent network connection (Mobile Games, 2001). Customers interact with mobile network operators as subscribers to the network service (Relationship 1 in Figure 1).

Mobile device supply manufacturers (e.g., *Nokia*, *Siemens*, *Motorola*) provide customers with the devices needed to connect to a network and to access the services provided. In the case of mobile gaming, devices might need extended functionality such as a very fast processor (Leavitt, 2003). Customers interact with devices as information technology (IT) end-users (Relationship 2 in Figure 1).

A feature of the mobile gaming industry sector is the coupling between device and network technology: a handheld device might be designed to be used only with a particular technology and for access to networks based on that technology. The IT end-user experience will depend on the characteristics of the device and the underlying wireless network technology.

In the interface category, some widely used middleware products for game development are WAP and iMode (Baldi & Thaung, 2002; Lee et al., 2003). New software platforms (J2ME and BREW) and operating systems (Symbian) allow for the development of portable mobile games. The industry participants in this category provide the service infrastructure needed by game developers and other related content providers, but typically do not interact directly with customers (although *Nokia* have developed a mobile phone microbrowser). Rather they serve as a link between the physical infrastructure and the business services that bring mobile gaming to the customer.

Table 1. Critical success factors, customer roles and industry players (extended from Petrova & Qu, 2006)

Critical Success Factor		Customer Role	Industry Participants
Facilitating conditions	Providing payment and billing mechanisms	**Consumer**	Aggregators/publishers/ mobile service infrastructure providers
Trialability	Providing opportunities for a free trial of a service		Aggregators/publishers
Compatibility	Meeting the gaming needs of specific customer segments	**Game player**	Game developers/ publishers
Observability	Emphasizing the social importance of playing a mobile game		Aggregators/ publishers
Image	Emphasising the status of the game player		Aggregators/ publishers
Normative beliefs	Building a critical mass to create social pressure		Aggregators/ publishers
Complexity	Ensuring easy to use gaming applications	**IT End-User**	Device manufacturers/ game developers
Trust	Alleviating security and privacy concerns		Mobile network operators/application service infrastructure providers
Relative advantage	Providing a ubiquitous and accessible mobile gaming service		Mobile network operators/application service infrastructure providers
Self-efficacy	Technical services matching different customer segment needs	**Subscriber**	Device manufacturers/ game developers

In the business services category, mobile game developers (for example *In-Fusio*) and mobile game aggregators and publishers (for example *MFORMA*, *Digital Bridges*) provide games to customers. Customers act as game players (Relationship 3 in Figure 1) and as consumers who pay for the valued added service (Relationship 4 in Figure 1).

CUSTOMER ROLES AND CRITICAL SUCCESS FACTORS

Understanding the priorities of a customer would lead to a better understanding of the criticality of the success factors of mobile gaming adoption and to more informed decision-making when offering a mobile gaming service to a targeted audience. From the discussion above, the question that arises is: What is the significance of customer roles in mobile gaming adoption?

Well known models and theories have been used to investigate customer adoption of different mobile applications (Barnes & Huff, 2003; Kleijnen et al., 2003; 2004; Pagani & Schipani, 2003; Carlsson et al., 2005; Pedersen, 2005; Yoon, Ha, & Choi, 2005). Key contributing factors based on customer attitudes and perspectives (Moore & Rutter, 2004; Pedersen, Methlie, & Thorbjornsen, 2002) and potential adoption drivers (Baldi & Thaung, 2002) have been identified. Petrova and Qu (2006) extracted a set of critical success factors for mobile gaming adoption based on findings from the literature, and proposed a framework for studying the adoption process (Table 1, first two columns). Extending this work further, in Table 1 (the last two columns) each factor is matched to a specific customer role based on the relationships discussed in the previous section.

Most of the critical success factors (seven) relate to two customer roles: "game player" and "IT end-user." They represent a relationship with the physical infrastructure layer and the business services layer of the mobile gaming industry, respectively, and constitute a "first priority" group of factors. The remaining three factors (matching the roles of "consumer" and of "subscriber") form the "second priority" group—also related to both infrastructure and business

Table 2. Future trends in mobile game development

Game Development Trends	Example
Games for 3G mobile networks	"Virtual Girlfriend" (*Artificial Life Inc.,* Hong Kong)
Multi-player, networked, interactive games	"MLSN Sports Picks" (*Digital Chocolate Inc.,* Finland)
Location–based games	"The Shroud" (*Your World Games,* U.S.)
Games for casual game players	Puzzle games (*Future Platforms,* UK). Backgammon (*Nokia* and *Octopi,* U.S.)
Highly personalised mobile gaming services.	"Vomitron" (*Ninemsn,* Australia). "PrizePlay" (*Sennari,* U.S.A). iMode in Japan, Europe and Australia.
3D games.	"V-Rally 3D" (*Fishlabs,* Germany)
Games which can be used for learning, or for awareness development	"FreedomHIV/AIDS" (*ZMQ,* India)

services provision. The significance of these results and their implications are discussed next.

Customer Priorities

Four factors are related to the role of the customer as a "game player." The implications are that mobile gaming will be successfully adopted by customers who are likely to enjoy mobile gaming as a social activity. Therefore mobile game content needs to be tailored to match the needs of particular social groups; developers will need to provide content of greater diversity—for example, multi-player games, building a virtual world (Raghu, Ramesh, & Whinston, 2002) and location-based gaming (Finn, 2005). The importance of the social context is confirmed by the success of mobile gaming in Japan (Barnes & Huff, 2003) and in other Asian countries (Kymalainen, 2004, p. 131). Game developers, aggregators and publishers should continue to focus on market segments that display a strong, pre-existing disposition towards mobile gaming as innovation. According to Baldi and Thaung (2002) and Kleijnen et al. (2004), Internet usage is such a predictor. Non-core gamer markets might also be penetrated with a range of casual games.

Three success factors relate to the role of the customer as an IT end-user. The implication for industry participants involved in the provision of physical infrastructure is that the usability of handheld devices will continue to be important as it brings "compelling" value to the customer (Venkatesh, Ramesh, & Massey, 2003). Network operators might need to consider "opening up their networks" to facilitate the development and provisioning of interactive games, operable across provider networks (Smorodinsky, 2002), while game developers will need to develop more portable applications (Yuan, 2004). All industry participants will need to address users' concerns for privacy and security, possibly by adapting solutions developed for e-commerce (Yiliant-tila, 2004, p. 71).

In summary, the adoption factors related to the consumer roles of "game player" and "IT end-user" are the ones that will ultimately determine the success of mobile gaming application and services.

Customer Support

Two factors relate to the customer role of "consumer" and one to the role of "subscriber." The implication is that that mobile game customers will expect a transaction and trading environment similar to other commercial environments, where customer and technical support are readily available (Kleijnen, de Ruyter, & Wetzels, 2003, p. 213). Industry participants from both the infrastructure layer and the business services layer will need to explore different cooperation strategies to be able to meet these requirements and so gain sustainable competitive advantage (Feldmann, 2002).

In summary, while adoption factors related to the customer roles of "consumer" and "subscriber" are not crucial, they will have a strong influence on customer choice and loyalty.

FUTURE TRENDS

The game development effort is focused on providing games targeting different customer segments, games exploring new technologies and games enhancing personalization. A summary of the future trends in mobile game development is provided in Table 2 along with some examples from the industry to illustrate them.

An important global industry trend is the move to define mobile game interoperability standards, so that game development can focus on producing games deployable across multiple game servers, wireless networks, and handheld devices. An example is the work of the Open Mobile Alliance (OMA) Games Services working group.

Academic work will continue to focus on multi-platform development and middleware architecture, on factors influencing adoption and on the development of innovative business models, studying customers from diverse cultural backgrounds and their needs and priorities.

CONCLUSION

The article reviews the literature on mobile games provision and adoption and outlines the different roles played by customers involved in mobile gaming. A mobile gaming framework is introduced, including critical success factors, customer roles and industry participants. Two priority groups of success factors are identified. The factors related to customers acting as "game players" and "IT end-users" are crucial for selecting and penetrating a market segment that is ready to adopt mobile gaming as a social activity; the factors related to customers acting as "consumers" and "subscribers" play an important but secondary role and might confer competitive advantage in the selected market. The future trends in mobile gaming will be towards the development of multi-player real-time games, and providing interoperability across platforms and networks, which will facilitate new market segment penetration.

REFERENCES

Anckar, B., & D'Incau, D. (2002). Value-added services in mobile commerce: An analytical framework and empirical findings from a national consumer survey. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 1087-1096).

Baldi, S., & Thaung, H. P-P. (2002). The entertaining way to m-commerce: Japan's approach to the mobile Internet—A model for Europe? *Electronic Markets, 12*(1), 6-13.

Barnes, S. J. (2003). The mobile commerce value chain in consumer markets. In *m-Business: The strategic implications of wireless technologies* (1st ed.) (pp. 13-37). Burlington MA: Butterworth-Heinemann.

Barnes, S. J., & Huff, S. L. (2003). Rising sun: iMode and the wireless Internet. *Communications of the ACM, 46*(11), 79-84.

Buellingen, F., & Woerter, M. (2004). Development perspectives, firm strategies and applications in mobile commerce. *Journal of Business Research, 57*(12), 1402-1408.

Carlsson, C., Hyvonen, K., Repo, P., & Walden, P. (2005). Asynchronous adoption patterns of mobile services. In *Proceedings of the 38th Annual Hawaii International Conference on Systems Sciences* (p. 189a).

Feldmann, V. (2002). Competitive strategy for media companies in the mobile Internet. *Schmalenbach Business Review*, *54*, 351-371. Retrieved January 5, 2006, from http://www.vhb.de/sbr/pdfarchive/einzelne_pdf/sbr_2002_oct-351-371.pdf

Finn, M. (2005). Gaming goes mobile: Issues and Implications. *Australian Journal of Emerging Technologies and Society*, *39*(1), 32-42

Funk, J. L. (2003). *Mobile disruption: The technologies and applications driving the mobile Internet*. New Jersey: John Wiley & Sons.

Graft, K. (2006, January 22). Analysis: A history of cellphone gaming. *BusinessWeek Online*. Retrieved January 25, 2006, from http://www.businessweek.com/innovate/content/jan2006/id20060122_077129.htm

Hagleitner, M., & Mueck, T. A. (2002). WAP-G: A case study in mobile entertainment. In *Proceedings of the 35th Hawaii International Conference on System Sciences* (Vol. 3) (p. 88).

Harmer, J. A. (2003). Mobile multimedia services. *BT Technology Journal*, *21*(3), 169-180.

Kleijnen, M. D., de Ruyter, K., & Wetzels, M. G. M. (2003). Factors influencing the adoption of mobile gaming services. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile Commerce: Technology, Theory, and Applications* (pp. 202-217). Hershey, PA: Idea Group Publishing.

Kleijnen, M. D., de Ruyter, K., & Wetzels, M. G. M. (2004). Customer adoption of wireless services: Discovering the rules, while playing the game. *Journal of Interactive Marketing*, *18*(2), 51-60.

Kymalainen, P. (Ed.). (2004). Mobile entertainment industry and culture. *European Commission User-Friendly Information Society.* Retrieved January 10, 2006, from http://www.mgain.org/mgain-wp8-D823_book-delivered.pdf

Leavitt, N. (2003). Will wireless gaming be a winner? *Computer, 36*(1), 24-27.

Lee, C.-W., Hu, W.-C, & Yeh, J.-H. (2003). A system model for mobile commerce. In *23rd International Conference on Distributed Computing Systems Workshops* (pp. 634-639).

Maitland, C. F., van de Kar, E. A. M., de Montalvo, U. W., & Bouwman, H. (2005). Mobile information and entertainment services: Business models and service networks. *International Journal of Management and Decision Making, 6*(1), 47-64.

Mobile Games (2001, June). *Mobile games*. *Mobile Entertainment,* pp. 4-50.

Moore, K., & Rutter, J. (2004). Understanding consumers' understanding of mobile entertainment. In K. Moore & J. Rutter (Eds.), *Mobile Entertainment: User-Centred Perspectives* (pp. 49-65). Manchester, UK: University of Manchester.

Natsuno, T. (2003). *I-mode strategy*. Chichester, UK: Wiley & Sons.

Ollila, M., Kronzell, M., Bakos, N., & Weisner, F. (2003). *Mobile entertainment business*. European Commission User-Friendly Information Society. Retrieved December 23, 2005, from http://www.mgain.org/mgain-wp5-D542-delivered3.pdf

Paavilainen, J. (2002). Consumer mobile commerce. In *Mobile Business Strategies*: *Understanding the Technologies and Opportunities* (pp. 69-121). London: IT Press.

Pagani, M., & Schipani, D. (2003). Motivations and barriers to the adoption of 3G mobile multimedia services: An end user perspective in the Italian market. In M. Khosrow-Pour (Ed.), *Proceedings of the 2003 Information Resources Management Association International Conference* (pp. 957-960). Hershey, PA: IRM Press.

Pedersen, P. E., Methlie, L. B., & Thorbjornsen, H. (2002). Understanding mobile commerce end-user adoption: A triangulation perspective and suggestions for an exploratory service evaluation framework. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 1079-1086).

Pedersen, P. E. (2005). Adoption of mobile Internet services: An exploratory study of mobile commerce early adopters. *Journal of Organizational Computing and Electronic Commerce*, *15*(3), 203-221.

Petrova, K. (2005). A study of the adoption of mobile commerce applications and of emerging viable business models. In M. Khosrow-Pour (Ed.), *Managing Modern Organizations with Information Technology. Proceedings of the 2005 Information Resources Management Association International Conference* (pp. 1133-1135). Hershey, PA: IRM Press.

Petrova, K., & Qu. H. (2006). Mobile gaming: A reference model and critical success factors. In M. Khosrow-Pour (Ed.), *Emerging Trends and Challenges in Information Technology Mangement: Proceedings of the 2006 Information Resources Mangement Association International Conference* (pp. 228-231). Hershey, PA: IRM Press.

Raghu, T. S., Ramesh, R., & Whinston, A. B. (2002). Next steps for mobile entertainment portals. *Computer*, *35*(5), 63-70

Shchiglik, C., Barnes, S., Scornavacca, E., & Tate, M. (2004). Mobile entertainment service in New Zealand: An examination of consumer perceptions towards games delivered via the wireless application protocol. *International Journal of Services and Standard*s, *1*(2), 155-171.

Siau, K., Lim, E., & Shen, Z. (2001). Mobile commerce: Promises, challenges and research agenda. *Journal of Database Management, 12*(3), 4-13.

Smorodinsky, R. (2002) Harnessing the potential of mobile entertainment. In *Digital Infrastructure Technology* (pp. 55-56). Retrieved January 5, 2006, from http://www.m-e-f.org/pdf/Harnessing%20the%20Potential.pdf

Stanoevska-Slabeva, K. (2003). Towards a reference model for m-commerce applications. In *Proceeding of the 2003 European Conference on Information Systems*. Retrieved March 1, 2004, from http://inforge.unil.ch/yp/Terminodes/papers/03ECISSG.pdf

Van de Kar, E., Maitland, C. F., de Montalvo, U. W., & Bouwman, H. (2003). Design guidelines for mobile information and entertainment services: Based on the Radio 538 ringtunes I-mode service case study. In *Proceedings of the 5th International Conference on Electronic Commerce* (pp. 413-417).

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications*, *3*(7), 185-187.

Venkatesh, V., Ramesh, V., & Massey, A.P. (2003). Mobile commerce opportunities and challenges: Understanding usability in mobile commerce. *Communications of the ACM, 46*(12), 30-32.

Wiener, S. N. (2003, August). *Terminology of mobile entertainment*. Mobile Entertainment Forum. Retrieved November 25, 2005, from http://www.m-e-f.org/pdf/GlossaryRelease_new%20logo.pdf

Yilianttila, M. (2004). *Emerging and future mobile entertainment technologies*. European Commission User-Friendly Information Society. Retrieved January 10, 2006, from http://www.mgain.org/mGain-wp4-d421-delivered-revised.pdf

Yoon, Y. S., Ha, I. S., & Choi, M.-K. (2005). Nature of potential mobile gamers' behaviour under future wireless mobile environment. In *Proceedings of the 7th International Conference on Advanced Communication Technology* (pp. 551-558).

Yuan, M. (2004, September 28). *Challenges and opportunities in mobile games*. IBM. Retrieved December 23, 2005, from http://www-128.ibm.com/developerworks/wireless/library/wi-austingameconf.html

KEY TERMS

Aggregator: A content provider who maintains relationships with several mobile network operators and delivers mobile game content to their users in a transparent way (i.e., using the same mobile number). Example: *MFORMA* (U.S.).

Binary Runtime Environment for Wireless (BREW): An application development platform for client applications designed for mobile networks using CDMA, based on the "C++" programming language.

Code Division Multiple Access (CDMA): A wireless technology used in 2G mobile networks, which is also developed for 3G services (e.g., W-CDMA, CDMA2002)

Developer: Typically a team of designers, artists, and engineers designing and creating game content (including graphics and sound). Example: *Future Platforms* (UK).

General Packet Radio Services (GPRS): An advanced technology for mobile data transmission based on GSM. Supports mobile multimedia applications and Web interaction, and enables pay-per-data models.

Global System for Mobile Communication (GSM): A standard for digital mobile telephony especially popular in Europe, but used worldwide. Provides across-border compatibility and maintains services at the 2G level.

Java 2 Platform Micro Edition (J2ME): An application development platform used primarily for client applications designed for mobile networks using GSM, based on the "Java" programming language.

Publisher: A company involved in controlling and marketing a mobile game, often funding its development. Example: *Your World Games* (U.S.).

Second Generation (2G) Mobile Telephony: Includes a range of technologies for wireless communication such as GSM and CDMA. Provides digitised voice communication, short messaging service (SMS) and some special features.

Symbian: An operating system for mobile devices with data processing capabilities (an open industry standard).

Third Generation (3G): "third generation." A range of advanced wireless communication technologies capable of supporting multimedia, video streaming, and video-conferencing.

Wireless Application Protocol (WAP): A set of open standards to enable Web and e-mail access and data display on handheld devices.

Mobile Healthcare Communication Infrastructure Networks

Phillip Olla
Madonna University, USA

INTRODUCTION

M-health is defined as “mobile computing, medical sensor, and communications technologies for healthcare” (Istepanian & Zhang, 2004). The use of the m-health terminology relates to applications and systems such as telemedicine and biomedical sensing systems (Budinger, 2003). The rapid advances in information and communication technology (ICT) (Godoe, 2000), nanotechnology, bio-monitoring (Budinger, 2003), mobile networks (Olla, 2005a), pervasive computing (Akyildiz & Rudin, 2001), wearable systems, and drug delivery approaches (Amy & Richards, 2004) are transforming the healthcare sector. The insurgence of innovative technology into healthcare practice is not only blurring the boundaries of the various technologies and fields, but is also causing a paradigm shift that is blurring the boundaries between public health, acute care, and preventative health (Hatcher & Heetebry, 2004). These developments have not only had a significant impact on current e-health and telemedical systems (Istepanian & Zhang, 2004), but they are also leading to the creation of a new generation of m-health systems with convergence of devices, technologies, and networks at the forefront of the innovation.

The phenomenon to provide care remotely using ICT can be placed into a number of areas such as m-health, telemedicine, and e-health. Over the evolution of telemedicine, new terminologies have been created, as new health applications and delivery options became available and the application areas extended to most healthcare domains. This resulted in confusion, and identification of what falls under telemedicine and what falls under telehealth or e-health became more complicated as the field advanced. New concepts such as pervasive health and m-health are also adding to this confusion. The first section of this article provides the background of telemedicine and the advancements of mobile networks, which are collectively the foundation of m-health. The evolution and growth of telemedicine is highly correlated with ICT advancements and software development. Telemedicine advancements can be categorized into three eras (Bashshur, Reardon, & Shannon, 2000; Tulu & Chatterjee, 2005) discussed in the next section.

There are numerous wireless infrastructures available for healthcare providers to choose from. Mobile networks that provide connectivity within buildings use different protocols from the standard digital mobile technologies such as global mobile systems (GSMs), which provide wide area connectivity. The second section of this article provides a summary of these mobile technologies that are having a profound impact on the healthcare sector. This section is then followed by the conclusion.

ERAS OF TELEMEDICINE

The *first era* of telemedicine solely focused on the medical care as the only function of telemedicine. This era can be named the telecommunications era of the 1970s. The applications in this era were dependent on broadcast and television technologies in which telemedicine applications were not integrated with any other clinical data. The *second era* of telemedicine was a result of digitalization in telecommunications, and it grew during 1990s. The transmission of data was supported by various communication mediums ranging from telephone lines to integrated service digital network (ISDN) lines. During this period there was a high costs attached to the communication mediums that provided higher bandwidth. The bandwidth issue became a significant bottleneck for telemedicine in this era. Resolving the bandwidth constraints has been a critical research challenge for the past decade, with new approaches and opportunities created by the Internet revolution; now more complex and ubiquitous networks are supporting the telemedicine. The *third era* of telemedicine was supported by the networking technology that was cheaper and accessible to an increasing user population. The improved speed and quality offered by Internet2 is providing new opportunities in telemedicine. In this new era of telemedicine, the focus shifted from an technology assessment to a deeper appreciation of the functional relationships between telemedicine technology and the outcomes of cost, quality, and access.

This article proposes a *fourth era,* which is characterized by the use of Internet protocol (IP) technologies, ubiquitous networks, and mobile/wireless networking capabilities, and can be observed by the proliferation of m-health applications that perform both clinical and non-clinical functions. Since the proliferation of mobile networks, telemedicine has attracted a lot more interest from both academic researchers and industry (Tachakra, Wang, Istepanian, & Song, 2003). This has resulted

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

in many mobile/wireless telemedicine applications being developed and implemented. Critical healthcare information regularly travels with patients and clinicians, and therefore the need for information to become securely and accurately available over mobile telecommunication networks is key to reliable patient care and reliable medical systems.

The telecommunication industry has progressed significantly over the last decade. There has been significant innovation in digital mobile technologies. The mobile telecommunication industry has advanced through three generations of systems and is currently on the verge of designing the fourth generation of systems (Olla, 2005b). The recent developments in digital mobile technologies are reflected in the fast-growing commercial domain of mobile telemedical services. Specific examples include mobile ECG transmissions, video images and tele-radiology, wireless ambulance services to predict emergency and stroke morbidity, and other integrated mobile telemedical monitoring systems (Istepanian & Zhang, 2004; New Scientist, 2005; Istepanian & Lacal, 2003; Warren, 2003). There is no doubt that mobile networks can introduce additional security concerns to the healthcare sector.

As security is a major concern, it is important to implement a mobile trust model that will ensure that a mobile transaction safely navigates multiple technologies and devices without compromising the data or the healthcare systems. M-health transactions can be made secure by adopting practices that extend beyond the security of the wireless network used and implementing a trusted model for secure end-to-end mobile transactions. The mobile trust model proposed by Wickramasinghe and Misra (2005) utilizes both technology and adequate operational practices to achieve a secure end-to-end mobile transaction. The first level highlights the application of technologies to secure elements of a mobile transaction. The next level of the model shows the operational policies and procedures needed to complement technologies used. No additional activity is proposed for the mobile network infrastructure since this element is not within the control of the provider or the hospital.

The next section will discuss the mobile network technologies and infrastructure which are key components of any m-health system; the network infrastructure acts as a channel for data transmission and is subject to the same vulnerabilities, such as sniffing, as in the case of fixed network transaction. The mobile networks discussed in the next section are creating the growth and increased adoption of m-health applications in the healthcare sector.

MOBILE HEALTHCARE COMMUNICATION INFRASTRUCTURE

The implementation of an m-health application in the healthcare environment leads to the creation of a mobile healthcare delivery system (MHDS). An MHDS can be defined as the carrying out of healthcare-related activities using mobile devices such as a wireless tablet computer, personal digital assistant (PDA), or a wireless-enabled computer. An activity occurs when authorized healthcare personnel access the clinical or administrative systems of a healthcare institution using mobile devices (Wickramasinghe & Misra, 2005). The transaction is said to be complete when medical personnel decide to access medical records (patient or administrative) via a mobile network to either browse or update the record.

Over the past decade there has been an increase in the use of new mobile technologies in healthcare such as Bluetooth and wireless local area networks (WLANs) that use different protocols from the standard digital mobile technologies such as 2G, 2.5, and 3G technologies. A summary of these technologies are presented below, and an overview of the speeds and range is presented in Table 1.

These mobile networks are being deployed to allow physicians and nurses easy access to patient records while on rounds, to add observations to the central databases, and to check on medications, among a growing number of other functions. The ease of access that wireless networks offer is matched by the security and privacy challenges presented by the networks. This serious issue requires further investigation and research to identify the real threats for the various types of networks in the healthcare domain.

Second-Generation (2G/2.5G) Systems

The second-generation cellular systems were the first to apply digital transmission technologies such as time division multiple access (TDMA) for voice and data communication. The data transfer rate was on the order of tens of kbit/s. Other examples of technologies in 2G systems include frequency division multiple access (FDMA) and code division multiple access (CDMA).

The 2G networks deliver high-quality and secure mobile voice and basic data services such as fax and text messaging, along with full roaming capabilities around the world. 2G technology is in use by more than 10% of the world's population, and it is estimated that 1.3 billion customers across more than 200 countries and territories around the world use this technology (GSM, 2005). The later advanced technological applications are called 2.5G technologies and include networks such as general packet radio service (GPRS) and EDGE. GPRS-enabled networks provide functionality such as: 'always-on', higher capacity, Internet-based content and packet-based data services enabling services such as color Internet browsing, e-mail on the move, visual communications, multimedia messages, and location-based services. Another complimentary 2.5G service is enhanced data rates for GSM evolution (EDGE), which offers similar capabilities to the GPRS network.

Table 1. Mobile networks

Networks	Speed	Range and Coverage	Main Issues for M-Health
2nd-Generation GSM	9.6 kilobits per second (KBPS)	World wide coverage, dependent on network operators' roaming agreements	Bandwidth limitation, interference
High-Speed Circuit-Switched Data (HSCSD)	Between 28.8 KBPS and 57.6 KBPS	Not global, only supported by service providers network.	Not widely available, scarcity of devices
General Packet Radio Service (GPRS)	171.2 KBPS	Not global, only supported by service providers network	Not widely available
EDGE	384 KBPS	Not global, only supported by service providers network	Not widely available, scarcity of devices
UMTS	144 KBPS—2 MBPS depending on mobility	When fully implemented, should offer interoperability between networks, global coverage	Device battery life, operational costs
Wireless Local Area	54 MBPS	30-50 m indoors and 100-500 m outdoors; must be in the vicinity of hot spot	Privacy, security
Personal Area Networks—Bluetooth	400 KBPS symmetrically, 150-700 KBPS asymmetrically	10-100m	Privacy, security, low bandwidth
Personal Area Networks—ZigBee	20 kb/s-250 KBPS depending on band	30m	Security, privacy, low bandwidth
WiMAX	Up to 70MBPS	Approx. 40m from base station	Currently no devices and networks cards
RFID	100 KBPS	1 m; non-line-of-sight and contactless transfer of data between a tag and reader	Security, privacy
Satellite Networks	400 to 512 KBPS new satellites have potential of 155 MBPS	Global coverage	Data costs, shortage of devices with roaming capabilities; bandwidth limitations

Third-Generation (3G) Systems

The most promising period is the advent of 3G networks, which are also referred to as the Universal Mobile Telecommunications System (UMTS). A significant feature of 3G technology is its ability to unify existing cellular standards, such as code-division multiple-access (CDMA), global system for mobile communications (GSM, 2005), and time-division multiple-access (TDMA), under one umbrella. Over 85% of the world's network operators have chosen 3G as the underlying technology platform to deliver their third-generation services (GSM, 2004). Efforts are underway to integrate the many diverse mobile environments in addition to blurring the distinction between the fixed and mobile networks. The continual roll out of advanced wireless communication and mobile network technologies will be the major driving force for future developments in m-health systems (Istepanian & Zhang, 2004). Currently the GSM version of 3G alone saw the addition of more than 13.5 million users, representing an annual growth rate of more than 500% in 2004. As of December 2004, 60 operators in 30 countries were offering 3GSM services. The global 3GSM customer base is approaching 20 million and has already been commercially launched in Africa, the Americas, Asia Pacific, Europe, and the Middle East (GSM, 2005), thus making this technology ideal for developing affordable global m-health systems.

Fourth Generation (4G)

The benefits of fourth-generation network technology include (Istepanian, Laxminarayan, & Pattichis, 2005; Olla, 2005a; Qiu, Zhu, & Zhang, 2002): voice-data integration, support for mobile and fixed networking, and enhanced services through the use of simple networks with intelligent terminal devices. 4G also incorporates a flexible method of payment for network connectivity that will support a large number of network operators in a highly competitive environment. Over the last decade, the Internet has been dominated by non-real-time, person-to machine communications (UMTS, 2002). The current developments in progress will incorporate real-time person-to-person communications, including high-quality voice and video telecommunications, along with extensive use of machine-to-machine interactions to simplify and enhance the user experience.

Currently the Internet is used solely to interconnect computer networks. IP compatibility is being added to many types of devices such as set-top boxes to automotive and home electronics. The large-scale deployment of IP-based networks will reduce the acquisition costs of the associated devices. The future vision is to integrate mobile voice communications and Internet technologies, bringing the control and multiplicity of Internet application services to mobile users (Olla, 2005b). 4G advances will provide both mobile patients and citizens the choices that will fit their lifestyle and make it easier for them to interactively get the medical attention and advice they need, when and where it is required, and how they want it, regardless of any geographical barriers or mobility constraints.

Worldwide Interoperability for Microwave Access (WiMAX)

WiMAX is considered to be the next generation of wireless fidelity (WiFi), wireless networking technology that will connect you to the Internet at faster speeds and from much longer ranges than current wireless technology allows (http://wimaxxed.com/). WiMax has been undergoing testing and is expected to launch commercially by 2007. The research firm Allied Business Research predicts that by 2009, sales of WiMax accessories will top US$1 billion (Taylor & Kendall, 2005), and Strategy Analytics predicts a market of more than 20 million WiMAX subscriber terminals and base stations per year in 2009 (ABI, 2005).

The technology holds a lot of potential for m-health applications, with the capabilities of providing data rates of up to 70 mbps over distances of up to 50 km. The benefits to both developing and developed nations are immense. There has been a gradual increase in popularity of this technology. Intel recently announced plans to mass produce and release processors aimed to power WiMax-enabled devices (WiMax, 2005). Other technology organizations investing to further the advancement of this technology include Qwest, British Telecom, Siemens, and Texas Instruments. They aim to get the prices of the devices powered by WiMax to affordable levels so that the public can adopt them in large numbers, making it the next global wireless standard. There are already Internet service providers in metropolitan areas offering pre-WiMAX service to enterprises in a number of cities including New York, Boston, and Los Angeles (WiMax, 2005).

.

Wireless Local Area Networks

Wireless local area networks (WLANs) use radio or infrared waves and spread spectrum technology to enable communication between devices in a limited area. WLAN allows users to access a data network at high speeds of up to 54 Mb/s as long as users are located within a relatively short range (typically 30-50 meters indoors and 100-500 meters outdoors) of a WLAN base station (or antenna). Devices may roam freely within the coverage areas created by wireless "access points," the receivers and transmitters connected to the enterprise network. WLANs are a good solution for healthcare today, plus they are significantly less expensive to operate than wireless WAN solutions such as 3G (Daou-Systems, 2001).

Personal Area Networks

A wireless personal area network (WPAN) (IBM Research, 2006; Istepanian & Zhang, 2004) is the interconnection of information technology devices within the range of an individual person, typically within a range of 10 meters. For example, a person traveling with a laptop, a PDA, and a portable printer could wirelessly interconnect all the devices, using some form of wireless technology. WPANs are defined by IEEE Standard 802.15 (IEEE Working Group, 2006). The most relevant enabling technologies for m-health systems are Bluetooth (http://www.bluetooth.org/) and ZigBee (http://www.zigbee.org/). ZigBee is a set of high-level communication protocols designed to use small, low-power digital radios based on the IEEE 802.15.4 standard for wireless personal area networks. ZigBee is aimed at applications with low data rates and low power consumption. ZigBee's current focus is to define a general-purpose, inexpensive, self-organizing network that can be shared by industrial controls, medical devices, smoke and intruder alarms, building automation, and home automation. The network is designed to use very small amounts of power, so that individual devices might run for a year or two with a single alkaline battery, which is ideal for use in small medical devices and sensors. The Bluetooth specification was first developed by Ericsson and later formalized by the Bluetooth Special Interest Group established by Sony Ericsson, IBM, Intel, Toshiba, and Nokia, and later joined by many other companies. A Bluetooth WPAN is also called a *piconet* and is composed of up to eight active devices in a master-slave relationship. A piconet typically has a range of 10 meters, although ranges of up to 100 meters can be reached under ideal circumstances. Implementations with Bluetooth versions 1.1 and 1.2 reach speeds of 723.1 kbit/s. Version 2.0 implementations feature Bluetooth Enhanced Data Rate (EDR) and thus reach 2.1 Mbit/s (http://www.bluetooth.org/; http://en.wikipedia.org/wiki).

Radio Frequency Identification (RFID)

RFID systems consist of two key elements: a tag and a reader/writer unit capable of transferring data to and from the tag. An antennae linked to each element allows power to be transferred between the reader/writer and remotely sited tag through inductive coupling. Since this is a bi-directional process, modulation of the tag antenna will be reflected back to the reader's/writer's antenna, allowing data to be transferred in both directions. Some of the advantages of RFID that makes this technology appealing to the healthcare sector are:

- no line-of-sight required between tag and reader;
- non-contact transfer of data between a tag and reader;
- tags are passive, which means no power source is required for the tag component;
- data transfer range of up to 1 meter is possible; and
- rapid data transfer rates of up to 100 Kbits/sec.

The use of RFID in the healthcare environment is set to rise and is currently being used for drug tracking. RFID technology is expected to decrease counterfeit medicines and make obtaining drugs all the more difficult for addicts (Weil, 2005). There are also applications that allow tagging of patients, beds, and expensive hospital equipment.

Satellite Technologies

Satellite broadband uses a satellite to connect customers to the Internet. Two-way satellite broadband uses a satellite link to both send and receive data. Typical download speeds are 400 to 512 kbps, while upload speeds on two-way services are typically 64 to 128 kbps. Various organizations (Inmarsat Swift64, 2000) have been investigating the development of an ultra-high-data-rate Internet test satellite for use for making a high-speed Internet society a reality (JAXA, 2005). Satellite-based telemedicine service will allow a real-time transmission of electronic medical records and medical information anywhere on earth. This will make it possible for doctors to diagnose emergency patients even from remote areas, and also will increase the chances of saving lives by receiving early information as ambulance data rates of 155mbps are expected. One considerable drawback associated with using this technology is cost (Olla, 2004).

CONCLUSION

The first section of this article provides the background of telemedicine and the advancements of mobile networks, which is fuelling the increase of m-health applications in the healthcare domain. The evolution and growth of telemedicine is highly correlated with ICT advancements and software development. This article has also summarized the various mobile network technologies that are being used in the healthcare sector. The mobile technologies described above have a significant impact on the ability to deploy mobile healthcare applications and systems.

REFERENCES

ABI. (2005). WIFI-WIMAX. Retrieved from http://www.abiresearch.com/category/Wi-Fi_WiMAX

Akyildiz, I., & Rudin, H. (2001). Pervasive computing. *Computer Networks—The International Journal of Computer and Telecommunications Networking, 35*(4), 371-371.

Amy, C., & Richards G. (2004). A BioMEMS review: MEMS technology for physiologically integrated devices. *Proceedings of IEEE* (Vol. 92, no. 1, pp. 6-21).

Bashshur, R. L., Reardon, T. G., & Shannon, G. W. (2000). Telemedicine: A new health care delivery system. *Annual Review of Public Health, 21,* 613-637.

Budinger, T. F. (2003). Biomonitoring with wireless communications. *Annual Review of Biomedical Engineering, 5,* 412.

Daou-Systems. (2001). *Going mobile: From eHealth to mHealth.* Retrieved from http://www.daou.com/emerging/pdf/mHealth_White_Paper_April_2001.PDF

Godoe, H. (2000). Innovation regimes, R&D and radical innovations in telecommunications. *Research Policy, 29*(9), 1033-1046.

GSM. (2005). *Homepage.* Retrieved from http://www.gsmworld.com/index.shtml

GSM. (2004). *Information.* Retrieved from http://www.gsmworld.com/index.shtml

Hatcher, M., & Heetebry, I. (2004). Information technology in the future of health care. *Journal of Medical Systems Issue, 28*(6), 673-688.

IBM Research. (2006). Retrieved from http://www.research.ibm.com/topics/popups/smart/mobile/html/phow.html

IEEE Working Group. (2006). *IEEE 802.15 Working Group for WPAN.* Retrieved from http://www.ieee802.org/15/

Imrasat Swift64. (2000). *Inmarsat announces availability of the 64kbit/s mobile office in the sky.* Retrieved from http://www.inmarsat.com/swift64/press_1.htm

Istepanian, R. S. H., & Lacal, J. (2003). M-health systems: Future directions. *Proceedings of the 25th Annual IEEE International Conference on Engineering Medicine and Biology,* Cancun, Mexico.

Istepanian, R. S. H., Laxminarayan, S., & Pattichis, E. (2005). *M-health: Emerging mobile health systems*. New York.

Istepanian, R. S. H., & Zhang, Y. T. (2004). Guest editorial introduction to the special section on m-health: Beyond seamless mobility and global wireless health-care connectivity. *IEEE Transactions on Information Technology in Biomedicine, 8*(4).

JAXA. (2005). *Aerospace Exploration Agency*. Retrieved February, 2006, at www.jaxa.jp/missions/projects/sat/index_e.html

New Scientist. (2005). Future trends, convergence IS. *New Scientist,* (March 12), 53.

Olla, P. (2004). A convergent mobile infrastructure: Competition or Co-operation. *Journal of Computing and Information Technology: Special Issue on Information Systems: Healthcare and Mobile Computing, 12*(4), 309-322.

Olla, P. (2005a). Evolution of GSM network technology. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking.* Hershey, PA: Idea Group Reference.

Olla, P. (2005b). Incorporating commercial space technology into mobile services: Developing innovative business models. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities.* Hershey, PA: IRM Press.

Qiu, R. C., & Zhang, Y. Q. (2002, May 26-29). Third-generation and beyond (3.5G) wireless networks and its applications. *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCS*), Scottsdale, AZ.

Tachakra, S., Wang, X. H., Istepanian, R. S. H., & Song, Y. H. (2003). Mobile e-health: The unwired evolution of TelemedicineMobile. *Telemedicine Journal and E-Health, 9*(3), 247.

Taylor, C., & Kendall, P. (2005, June 1). *Strategy analytics: WiMAX 3G killer or fixed broadband wireless standard?* Retrieved from www.strategyanalytics.net/default.aspx?mod=ReportAbstractViewer&a0=2393

Tulu, B., & Chatterjee, S. (2005). A taxonomy of telemedicine efforts with respect to applications, infrastructure, delivery tools, type of setting and purpose. *Proceedings of the 38th Hawaii International Conference on System Sciences,* Hawaii.

UMTS. (2002). *Support of third generation services using UMTS in a converging network environment.* Retrieved from http://www.umts-forum.org/servlet/dycon/ztumts/umts/Live/en/umts/Resources_Reports_index: UMTS

Warren, S. (2003). Beyond telemedicine: Infrastructures for intelligent home care technology. *Proceedings of the Pre-ICADI Workshop Technology for Aging, Disability, and Independence,* London.

Weil, N. (2005). Companies announce RFID drug-tracking project: Unisys and SupplyScape plan to track Oxycontin through the supply chain. *Computerworld.*

Wickramasinghe, N., & Misra, S. K. (2005). A wireless trust model for healthcare. *International Journal of Electronic Healthcare, 1*(1), 62.

WiMax. (2005). Retrieved from http://wimaxxed.com

KEY TERMS

Bluetooth: Worldwide industrial specification for wireless personal area networks (PANs). Bluetooth provides a way to connect and exchange information between devices like personal digital assistants (PDAs), mobile phones, laptops, PCs, printers, and digital cameras via a secure, low-cost, globally available short range radio frequency.

Global System for Mobile Communications (GSM): A digital worldwide mobile phone and data standard. GSM service is used by over 1.5 billion people across more than 210 countries and territories. The ubiquity of the GSM standard makes international roaming very common between mobile phone operators, enabling subscribers to use their phones in many parts of the world.

M-Health: Mobile computing, medical sensor, and communications technologies for healthcare.

Personal Area Network (PAN): A computer network with a reach of a meters used for communication among computer devices such as telephones and personal digital assistants or medical sensors close to the human body.

Radio Frequency Identification (RFID): An automatic identification system that transmits, stores, and remotely retrieves data using mobile devices called RFID tags or transponders. The purpose of an RFID system is to enable data to be transmitted by a mobile device, called a tag, which is read by an RFID reader and processed according to the needs of a particular application.

Mobile Hunters

Jörg Lonthoff
Technische Universität Darmstadt, Germany

INTRODUCTION

Growing Internet mobility due to various transmission methods such as broadband data transmission get mobile service providers interested in providing services that offer more than voice telephony. Modern cellular phones support general packet radio service (GPRS) have a color display and are usually Java-compliant. This meets the device's requirements for context-based services. As global system for mobile communications (GSM)-based cellular phones are widely used and, at least in Europe, the GSM-network is available almost everywhere, the context variable "location" seems useful for extending the relevant value-added services (Rao & Minakakis, 2003, p. 61). For example, there are services for finding friends in the vicinity (Buddy Alert by Mobiloco, www.mobiloco.de) and mobile navigation systems for cellular phones (NaviGate by T-Mobile, www.t-mobile.de/navigate). But there has not yet been a real breakthrough for location-based services (LBSs) (Lonthoff & Ortner, 2006).

Supported by T-Systems International, we developed the adventure game "Mobile Hunters." The game demonstrates what is possible with LBS and uses the currently available infrastructure mobile network providers offer for creating a virtual playing field. This playing field will be adapted to the real world. The object of the game is a hunt. Players can either be a hunter who must find a fugitive or a fugitive who has to make sure he is not getting caught. Of course, this hunt will become eventful as there are a variety of obstacles. Playing a so-called mobile location-based game (MLBG) could increase the acceptance of further LBSs.

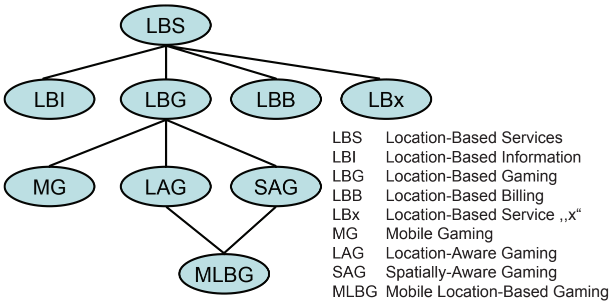
This article will first summarize the current state of research in this field and then present *mobile location-based gaming*. Then the reader will get to know the game "Mobile Hunters." After that, we will discuss the lessons learned from the game, and possible further developments will be considered. At the end of this, we draw a short conclusion.

BACKGROUND

MLBGs are a special category of location-based games, as follows:

> *A MLBG is a location-based game running on a mobile device. By using a communication channel the game exchanges information with a game server or other players.* (Lonthoff & Ortner, 2006)

Figure 1. Taxonomy for location-based services


LBS
LBI
LBG
LBB
LBx
MG
LAG
SAG
MLBG
LBS Location-Based Services
LBI Location-Based Information
LBG Location-Based Gaming
LBB Location-Based Billing
LBx Location-Based Service ,,x"
MG Mobile Gaming
LAG Location-Aware Gaming
SAG Spatially-Aware Gaming
MLBG Mobile Location-Based Gaming

Applying this definition, the fields *location-aware games* and *spatially aware games* become relevant. Figure 1 shows all terms relevant in MLBG.

Location-Based Services

The added value of mobile services opens up opportunities for service providers to address a new dimension of the user: the user's spatiotemporal position. Such services are called location-based services. LBSs are based on a variety of localization methods for determining a user's position. In the field of so-called *context-aware computing* (Schilit, Adams, & Want, 1994, p. 85), LBSs provide location information as context references (Dey, 2001). There are many possibilities of using the location reference in an application system (Unni & Harmon, 2003, p. 417; Schiller & Voisard, 2004). All of these services are based on mobile positioning. Mobile positioning comprises all technologies for determining the location of mobile devices. A position can be determined in two different ways: using network-based technology (the network provides the position) or using terminal-based technology (the device provides the position).

Network-Based Positioning Technologies

GSM-networks offer basically six different methods of network-based localization (Röttger-Gerigk, 2002). Cell of

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

origin (COO) is the simplest mobile positioning technique. It identifies the cell (cell ID) in which a cellular phone is logged on. The cell ID is connected with the radiation range of a mobile base station. The cellular coverage area has a certain range around the position of the mobile base station. One mobile base station can have several radiation areas (cell IDs), whereby these cells always refer to the same geographic position of the mobile base station's location (Hansmann, Merk, Nicklous, & Stober, 2001, p. 243). The positioning accuracy that can be achieved depends on the size of the cellular coverage area; it may range between 25 m and 35 km in diameter. In addition, there are more complex techniques such as angle of arrival (AOA), time of arrival (TOA), time difference of arrival (TDOA), signal attenuation (SA), and the radiocamera system.

Terminal-Based Positioning Technologies

Cell of origin can also be considered a terminal-based technique, as the desired cell ID can be read out directly from the device (terminal). To do this, however, a reference database is needed that contains the geographic coordinates stored for each cell ID. The following further terminal-based techniques are available: enhanced observed time difference (E-OTD), as well as the satellite-based systems such as the global positioning system (GPS), or assisted-GPS (A-GPS), which works without modifications to the cellular phone network infrastructure, except that the mobile device must possess a GPS receiver.

Gaming

Mobile games for cellular phones are currently experiencing a growing demand. There is a trend towards more complicated 3D games. This trend is supported by the current hardware development that is the availability of high-performance cellular phones or smart phones, respectively. Games that allow direct communication with remote participants are of great interest (multi-player games). Multi-player games on offer can be played using a wireless application protocol (WAP) portal or locally by two people (infrared) or by several players (Bluetooth).

Games for personal digital assistants (PDAs) are also very interesting. Such games are usually intended for one player. But games for several players become possible, if infrared, Bluetooth, or wireless local area network (WLAN) are used.

In location-based games the movements of a player (in the sense of a geographical change of location) influence the game. Nicklas, Pfisterer, and Mitschang (2001, pp. 61-62) suggest a classification of location-based games into mobile games, location-aware games, and spatially aware games.

Mobile games require as a location reference only one more player who is in the vicinity. The location information itself is not considered in the game. A typical example of this kind of game is Snake, a game of dexterity for two delivered with the older Nokia cellular phone models that can be played using infrared or Bluetooth. Location-aware games include information about the location of a player in the game. A typical example would be a treasure quest whereby a player must reach a particular location. Spatially aware games adapt a real-world environment to the game. This creates a connection between the real world and the virtual world. The MLBG "Mobile Hunters" presented in the following belongs to this category of games.

CHARACTERISTICS AND CHALLENGES OF MLBG

Important for MLBG are the type of device used, the communication and network infrastructure it is based on, the way positions are determined, and the kind of game.

Devices such as cellular phones, smart phones, and PDAs can be used, possibly laptop also. In addition to this rough classification, the device properties can serve for further distinction: the operating system, client programming (Java virtual machine, Web-client/WAP-client), the types of available user interfaces, as well as battery life and processor power.

The relevant communication media are wide area networks such as WLAN, GSM, and universal mobile telecommunication system (UMTS). These technologies vary in range and bandwidth. The accuracy of a determined position depends on the technique used and on the network structure.

When looking at the type of game, two dimensions are of interest: the number of players and the type of game. There are single-player games and multi-player games. You can also play multi-player games alone, if players are simulated. Massive-multi-player games are a special type of game in which the end of the game is not defined. Players can actively participate in the game for some time and improve their ranking in the community associated with the game. Relevant genres of game would be role-playing games, scouting games, real-time strategy games, and first-person-shooter games.

Users can find a variety of game collections on the Internet that include MLBGs. For example:

- www.smartmobs.com/archive/2004/12/28/location-based_.html
- www.we-make-money-not-art.com/archives/001653.php
- www.in-duce.net/archives/locationbased_mobile_phone_games.php

Newer publications offer studies (Jegers & Wiberg, 2006) and overviews (Magerkurth, Cheok, Nilsen, & Mandryk, 2005; Rashid, Mullins, Coulton, & Edwards, 2006) of pervasive games, including MLBGs.

What is fascinating about MLBG is: "This ability for you to actually use your real-world movements to play the game means that you are no more playing a game…You are in the Game!" (Mikoishi, 2004).

If you want to turn a "classic" game into an MLBG, there are four central problem areas, identified by Nicklas et al. (2001, p. 62): adaptation of the playing field, adaptation of the pawn in a game, representation of cards or objects, and adaptation of the moves in the game.

Another challenge results from the characteristics of mobile networks. In MLBG it may happen that some of the players interrupt the connection for short periods of time. These interruptions must be covered.

MOBILE HUNTERS

The Game

The idea of the adventure game "Mobile Hunters" was inspired by the well-known board game "Scotland Yard" (Ravensburger, 1983). The creation of a players' community reflects partly the notion of massive multi-player games. To be able to start a game session, at least two players must participate. There are two possible roles in the game: one or more hunters and one fugitive. The hunters want to catch the fugitive before the specified playing time is over. The fugitive must try to escape the hunters or, to prove that he is innocent, collect a number of items as proof of his innocence.

After logging onto the "Mobile Hunters" server with user name and password as authentication, a player can initiate a game. Once a game has been initiated, several more players can enter into the game. If a new player enters into the game, the game server checks whether the potential player's location is within an appropriate distance to the center of the playing field. This ensures that the spatial distance between the players does not become too large. The person who initiated the game decides when the number of players is sufficient and then starts the game. The maximum number of players can be specified.

When the game is started, the game server randomly assigns the roles and informs the players about whether they are a hunter or a fugitive. When the game has been started, the game server distributes all the items for hunters and fugitives randomly on the playing field. After this initialization phase the game begins synchronously on all participating clients and the countdown for the playing time starts. In previously specified intervals (default 1.5 minutes), the current position of the fugitive appears on a map in the hunters' display. In the playing field there are a number of locked (virtual) boxes that players can open if their position is the same as the position of the box (geo-coordinate with specified radius = playing field). Some of the boxes are visible only for hunters. These boxes contain offensive weapons. The fugitive can only see the boxes that contain items for him. These are items for defense or proof of innocence. Table 1 gives an overview of the various items.

A player can attack, if one player is located in the same position and the attacker has a weapon. If a hunter attacks another hunter, the person attacked will become incapable of action for some time (default 30 seconds)-his client's menu will be hidden. In this attack the attacker loses his weapon. If a hunter attacks a fugitive, the fugitive can defend himself using a matching item for defense. If this happens, the attacker will become incapable of action for some time (default 30 seconds). If the fugitive cannot defend himself, the hunter wins.

The fugitive wins if he is able to find a certain number (default 3) of items that prove his innocence or if no attack

Table 1. Overview of the items available for the players

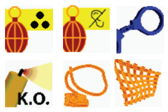
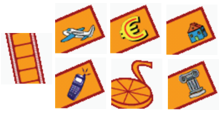
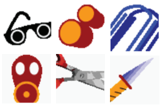
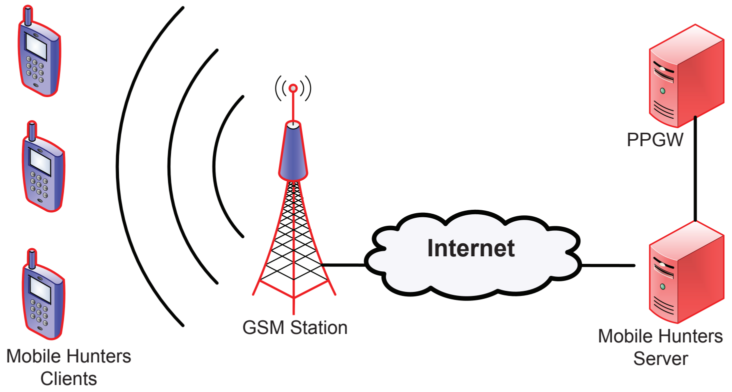
Recipient	Type	Instance	Purpose
Hunter	Attack		In the boxes the hunter will find offensive weapons he can use to attack the fugitive or for attacking another hunter to make him incapable of action (e.g., handcuffs).
Fugitive	Proof		In the boxes the fugitive will find items that can prove his innocence (e.g., a theater ticket).
Fugitive	Defense		In the boxes the fugitive will find items he can use to defend himself against attacks by a hunter (e.g., paper clip for defense against handcuffs)

Figure 2. Overall architecture



on him was successful during the duration (default 30 minutes) of the game.

Architecture

In the development of "Mobile Hunters," we only used technologies that have already been accepted in the market and are, or will most certainly be, widely spread, to facilitate acceptance of the game. We chose the GSM network, which is currently the most widely used (Nicklas et al., 2001, p. 61), to ensure a wide range of application. We chose a Java-compliant cellular phone (Nokia 6680) as the playing device.

It was essential that the positioning technique would only use the available cellular phone network infrastructure and that it would not be necessary to make modifications to it. This is why COO seemed best suited.

The implementation is based on a client/server architecture. Communication between the components is realized using the German T-Mobile GSM network. For data transfer we chose GPRS; communication takes place on an application level via secure hypertext transfer protocol (HTTPS). The "Mobile Hunters" server is a Java-based game server, which provides the user management and the game management. For determining the current position of each player, T-Systems' research platform "Permission and Privacy Gateway" (PPGW) is used. PPGW also provides the "Mobile Hunters" server with maps of the areas in question. Figure 2 shows the overall architecture schematically.

Client

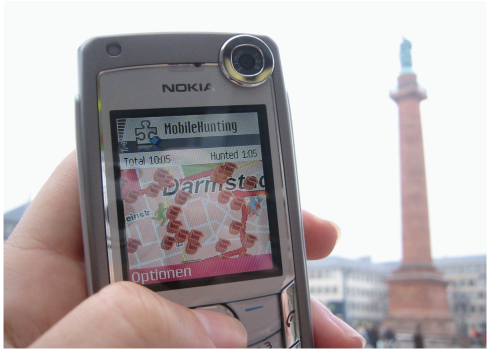
The "Mobile Hunters" client (see Figure 3) is the game's user interface. It displays the current map segments; here, interaction with the other players takes place. Players also use the "Mobile Hunters" client to register and log on.

Current prerequisites are a cellular phone with Symbian operating system version 7.0 and a Java 2 Micro Edition (J2ME) framework. A class implemented in Symbian C++ is needed for reading out the cell ID the cellular phone uses. The cell ID is transmitted to the "Mobile Hunters" server. The server requests the current status in specified intervals (default every 10 seconds). The necessary unique identification of each player is the user's name he or she entered at the beginning of the game.

Server

The "Mobile Hunters" server's task is to centrally control the game. On the server, identification and authentication of the users takes place, as well as the game logic and the communication between PPGW and the clients. Every initialization as well as the specification of the playing field happens dynamically on the game server. The server was implemented in Java as an Apache Tomcat application on a Microsoft Windows 2003 Server platform and connected

Figure 3. "Mobile Hunters" client on Nokia 6680 at Luisenplatz, Darmstadt, Germany



to a Microsoft SQL Server via Java database connectivity (JDBC)/open database connectivity (ODBC). This ensures a clear separation of application level and data level. The server establishes the connection to the PPGW for requesting the geographic data of each cell ID used in the game at any time. The PPGW has a reference database where the relevant geo-reference-data for each cell ID are stored. The geographic data that correspond to the cell IDs are returned by the PPGW via an eXtensible Markup Language (XML)-based interface. The server keeps a high-score list to challenge the gaming community. Using a configuration file, the parameters that can be controlled are passed to the server.

LESSONS LEARNED

The adaptation of the game's concept to the virtual world turned out to be difficult because of the inaccuracy of positioning. Cell switches occur even if a player does not move. This may happen, for example, if several participants are logged on to a cell and, due to the limits of bandwidth or for reasons of optimization, a neighboring cell is allocated to the cellular phone.

Cellular coverage areas have different sizes: in an urban area approximately 25 m, in rural areas up to 35 km. Therefore "Mobile Hunters" is suited for playing in high-density urban areas. The cellular coverage areas of single cells overlap with those of other cells. Large cells cover smaller cells almost completely. It is still under discussion whether to exclude cells from the game if they are too large and from which size on they should be banned. To increase the accuracy of positioning, GPS or A-GPS offers interesting opportunities for mobile network providers and for the MLBG.

When a "classic" game is adapted to an MLBG, it is important to make sure the point of the game is kept and that the duration of the game is adequate. The hype phases we observe are much shorter with games than with any other service offers. Once a game is considered boring or error prone, acceptance drops. For the game "Mobile Hunters" to be funny and exciting, a game duration of 30 minutes is recommended. The interval in which a fugitive's current position is displayed should be 1.5 minutes, the time a player is incapable of action should be 30 seconds, and the interval in which the cell ID is read ought to be less than one second. A powerful processor such as the Nokia 6680 (220 MHz) is necessary for playing this game. It is very important to develop the client's code in a way best suited for cellular phones (Lonthoff, Ortner, & Wolf, 2006). Session handling is essential for playing an MLBG. The user interface must be designed quite simply, so that no explanations will be necessary.

Our mobile-specific modeling resulted in a very economic consumption of resources and of handling communication data. In a 30-minute game, only an average of 300 kB data is transferred. Half of that data volume is used for loading the graphics (maps and icons) at the beginning of the game.

CONCLUSION

The conclusion we draw is that we have partly succeeded in adapting the real world to a game's virtual world. It is possible to read out the cell ID from a variety of mobile devices using different techniques. However, no standard application programming interface (API) is currently available to offer this functionality. This means that anyone who wants to create a game needs to develop a suitable API for every single cellular phone, if the game is to be widely used. This problem may be solved by deploying a middleware such as BREW from Qualcomm Inc. (www.qualcomm.com/brew; Tarumi, Matsubara, & Yano, 2004, p. 546).

Addressing the human play instinct, MLBGs increase the acceptance of LBSs. Playing, the use of LBSs will become effortless and people will get interested in such systems. Location-based games are becoming a mass market. If this mass market can be served, prices for location requests will go down.

The experience gained in the field of gaming also applies to situations in private and business life. Advertising and interactive marketing (Han, Cho, & Choi, 2005, p. 103) are potential application domains. Possibly, brand-new application systems useful in everyday life can be developed and implemented. In this context "application system" is understood in a comprehensive way for all tasks in user and computer-based information processing (Ortner, 2005, p. 34).

REFERENCES

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4-7.

Han, S., Cho, M., & Choi, M. (2005). Ubitem: A framework for interactive marketing in location-based gaming environment. *IEEE Proceedings of the International Conference on Mobile Business (ICMB'05)* (pp. 103-108), Sydney, Australia.

Hansmann, U., Merk, L., Nicklous, M., & Stober, T. (2001). *Pervasive computing handbook.* Berlin: Springer-Verlag.

Jegers, K., & Wiberg, M. (2006). Pervasive gaming in the everyday world. *Pervasive Computing, 5*(1), 78-85.

Lonthoff, J., & Ortner, E. (2006). Mobile location-based gaming (MLBG) as enabler for location-based services (LBS). *Proceedings of the E-Society,* Dublin, Ireland (pp. 485-492).

Lonthoff, J., Ortner, E., & Wolf, M. (2006). Implementierungsbericht Mobile Hunters. In J. Roth, J. Schiller, & A. Voisard (Hrsg.): *3. GI/ITG KuVS Fachgespräch Ortsbezogene Anwendungen und Dienste* (pp. 26-31), Technical Report, FU Berlin, Germany.

Magerkurth, C., Cheok, A.D., Nilsen, T., & Mandryk, R. (Eds.). (2005). *Proceedings of PerGames 2005,* Munich, Germany.

Mikoishi. (2004). *gunslingers.* Retrieved from http://www.gunslingers.mikoishi.com

Nicklas, D., Pfisterer, C., & Mitschang, B. (2001). Towards location-based games. *Proceedings of the International Conference on Applications and Development of Computer Games in the 21st Century (ADCOG 21)* (pp. 61-67), Hong Kong Special Administrative Region, China.

Ortner, E. (2005). *Sprachbasierte Informatk—Wie man mit Wörtern die Cyber-Welt bewegt.* Leipzig, Germany: Eagle-Verlag.

Ravensburger. (1983). *Scotland Yard.* Ravensburg, Germany.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Rashid, O., Mullins, I., Coulton, P., & Edwards, R. (2006). Extending cyberspace: Location based games using cellular phones. *ACM Computers in Entertainment (CIE), 4*(1), 1-18.

Röttger-Gerigk, S. (2002). Lokalisierungsmethoden. In *Handbuch mobile-commerce* (pp. 419-426). Berlin: Springer-Verlag.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *Proceedings of the Workshop on Mobile Computing Systems and Applications* (pp. 85-90), Santa Cruz, CA.

Schiller, J., & Voisard, A. (2004). *Location-based services.* San Francisco: Morgan Kaufmann.

Tarumi, H., Matsubara, K., & Yano, M. (2004). Implementations and evaluations of location-based virtual city system for mobile phones. *IEEE Proceedings of the Global Telecommunications Conference Workshops* (pp. 544-547), Dallas, TX.

Unni, R., & Harmon, R. (2003). Location-based services: Models for strategy development in m-commerce. *IEEE Proceedings of the Portland International Conference on Management of Engineering and Technology* (PICMET'03) (pp. 416-424), Portland, OR.

KEY TERMS

General Packet Radio Service (GPRS): An IP-based data communication service for cellular phones using GSM networks.

Global System for Mobile Communications (GSM): The most popular standard in the world for second-generation (2G) cellular phone systems.

Java 2 Micro Edition (J2ME): A special Java environment for small but smart devices, using a minimum of resources.

Location-Based Service (LBS): A context-based service using the context variable "location." In the area of m-commerce and m-business, LBSs enable new value-added services by using the customers' location.

Mobile Location-Based Game (MLBG): A MLBG is a location-based game running on a mobile device. By using a communication channel, the game exchanges information with a game server or other players.

Mobile Hunter: A pretty nice MLBG developed by the Department of Commercial Information Technology I at Technische Universität Darmstadt, Germany, in cooperation with T-Systems International.

Mobile Positioning: Summarizes technologies for providing the location of a mobile device.

Permission and Privacy Gateway (PPGW): A T-Systems research and development platform that provides a location interface. The PPGW was developed as a location brokerage platform that aggregates location sources (cell ID, WLAN, A-GPS, RFID) and offers this information to service providers using a unified interface.

Mobile ICT

Dermott McMeel
University of Edinburgh, Scotland

INTRODUCTION

Within building construction, the appropriation of a specific technology can vary across different job functions and participants (designer, foreman, stone mason, etc.). Embraced by some yet ignored by others, we can illicit insights into certain technologies and devices by reflecting not only on where they are successful but where they are problematic (Wiszniewski, Coyne, & Christopher, 1999) causing breakdowns and dysfunction in systems.

The rugged nature of the site condition quite often prohibits the proliferation of very delicate or expensive mobile equipment. Laser levels, point cloud generators, proprietary information and communication technology (ICT) devices, and so forth are limited to specific instances on the site where the usage is supervised and controlled, and all too often, “construction organisations have found that the ICT investment has failed to meet their expectations” (Peansupap & Walker, 2005). Robust, less expensive equipment is not subject to such extensive regulation (tape measures, bubble levels, nail guns), and their usage is often more prolific and creative. Various groups are charged with—and attempt to—fit ICTs into the current construction process; these ICTs are shown to create a much ‘smoother’ flowchart (COMIT, 2004). It is an established fact that the way in which a problem is presented generates a particular type of solution (Ortony, 1979; Schön, 1979); we would then suggest that contrary to popular representation and documentation of the construction process (Cox & Hamilton, 1995), there are perhaps more revealing models for presenting and understanding the communicative processes of the construction site than the established flowchart. Focusing on the ICT technology of mobile phones, we reflect on three different roles within a construction project: a director of a large construction organization, a site manager within that organization, and ‘micro-contractors’ of construction—that is, specialist subcontractors or small building contractors. We explore the usage and effects of the mobile phone in these different roles within a construction project.

OFFICIAL/UNOFFICIAL: THE FALSE DICOTOMY

There has always been unregulated communication within regulated work, and mobile phone usage has contributed to powerful unregulated and ‘unofficial’ (in the litigious sense) means for communication within construction. Theorists readily draw on concepts of space and containment to define communications. At the very least, language is an exercise in categorization, assuming similar meanings under a particular word or sign. Reddy (1979) suggests that these assumptions contribute to miscommunication, particularly when communicating across differing cultural categories. In formal communications there is an understanding that certain communiqués are meant for particular recipients, within certain categories of communications. This bureaucratization of communications has its place, but communication also requires the transgression of boundaries (Deleuze, 1988; Shannon & Weaver, 1963).

It has been suggested in a previous paper that parallels exist between construction and Carnival (McMeel, Coyne, & Lee, 2005). These unofficial means breakdown boundaries and thresholds (a distinctive feature of Carnival) and enables unexpected interaction, not unlike the ‘crossroads’ extensively employed when discussing native American Trickster figures (Hyde, 1998). These unexpected interactions cause discomfort and empowerment, both of which are symptoms of the Carnival condition.

CONSTRUCTION AS CARNIVAL

> *All the symbols of the Carnival idiom are filled with this pathos of change and renewal…of prevailing truths and authorities. We find here a characteristic logic, the peculiar logic of the ‘inside out’ (a l’envers), of the ‘turnabout,’ of a continual shifting from the top to bottom, from front to rear, of numerous parodies and travesties.* (Bakhtin, 1984)

The construction process too has “prevailing truths and authorities”; we suggest that mobile phones have contributed to further breakdown or “suspension, both ideal and real, of hierarchical rank” (Bakhtin, 1984) within construction. Other features of Carnival introduce notions of re-interpretation and redundancy (Attali, 1985). Here we provide theoretical grounding to such assertions, and draw upon our findings from interviews and observation on the construction site and discuss the pros and cons of mobile ICT in construction under the emergent themes of contiguity, abstractedness, porosity, and instantaneity.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Contiguity (Contact)

This temporary suspension, both ideal and real, of hierarchical rank created during Carnival time a special type of communication impossible in everyday life…permitting no distance between those who came in contact with each other and liberating from norms of etiquette and decency imposed at other times. (Bakhtin, 1984)

According to Bakhtin, the renewal and revitalization are the hallmarks of the Carnival, and are brought about when hierarchical barriers are momentarily dropped and populations cross-pollinate. This created an intriguing relationship between high and low society, and the interstices between. Ritual has played—and arguable continues to play—a part in construction (Jones, 2000). We have previously suggested (McMeel et al., 2005) that traditional ritualistic ceremonies (groundbreaking, topping of) served as a melting pot for laborers, architects, clients, and engineers—groups who would not otherwise meet—to mix. Pedreschi (2000) reflects on the architect/engineer Eladio Dieste and attributes his success, in part, to his skill in choosing excellent interlocutors to work for him, thus his team encouraged discussion rather than dictation.

Our findings suggest that mobile phones encourage this crossing of boundaries and reveal the predominant tendency by users to contact the 'top of the pile', as in the case of the director of a small specialist stair manufacturer subcontractor. If the director (SF) as the 'top of the pile' is contacted, it will indeed garner results, by virtue of the fact that his company is relatively small and SF oversees every set of stairs manufactured. Within a large construction company, contacting the director (EMC) who is 'out of the loop' of the day-to-day running of most projects would still allow him to re-direct the query to achieve resolution. Either way contiguity was expected and, when not achieved, often generated feelings of offense, as discussed with SF who deposits his mobile phone with the administration staff when he is attending a meeting or is on the workshop floor. The administrative staff regularly encounters hostility from callers who take exception to calling SF's personal number and speaking to someone else.

Distraction (Abstractedness)

Shock from the Carnival-esque can be caused by distraction from it; when surrounded by and embraced, it can be both surreal, exciting, and invigorating; if however it is thrust upon you, it is distracting and perhaps frightening. The construction site differs here in that distraction when in a dangerous environment can be fatal. In the words of EMC: "If you can't use a mobile phone when you're driving, you shouldn't be allowed to use one if your operating heavy machinery should you?"

Perhaps the process of 'making' things (either stairs or buildings) has its own unique environment, one of noise and dirt in this particular case. The 'making' process perhaps suffers from distraction, not from traditional notions of distraction (i.e., noise and dirt), but from the distraction of outside interventions: "If I'm on the workshop floor it's not to wander round and have a break…I'm probably attending to something a dam side more important than a phone call" (SF). Several other examples were discussed where site workers received calls (while guiding cranes, for example), and while no accidents ensued, their concentration on the task at hand was compromised. Whether it was the skill of the workshop environment or individuals in high-risk areas, distractions were problematic and repeatedly caused by incoming messages or calls.

Porosity (Screening)

This comes from the Greek 'pore', literally meaning 'passage' or 'gateway' (OED). When finding the gateway closed ,we find callers either lose interest, as in the case of the estate agents who would call CR (site manager of a large construction company) on a daily basis. The information they require can be obtained elsewhere; CR as the site manager is simply most up to date on matters such as completion time, floor areas, and so forth—he is the 'top of the (sub-)pile', so to speak. Or callers focus their attention on times when the 'gateway' is open, in the case of site manager CR at designated break times, when he would be in his site hut close to his phone.

When receiving calls we have seen smaller 'micro-contractors' (7-10 employees) prioritize their calls. By looking at the name and number, they can decide whether the call requires immediate attention or could wait or was unimportant and could be ignored. Depending on the decision, the call is answered, noted for a call back, or ignored.

Within the realm of Web portals, permeability and porosity has been explored (Coyne, Lee, & Parker, 2004) in terms of commerce, discourse, and interaction. At an organizational site level, the porosity of the mobile phone allowed for incoming calls to be prioritized by knowing who was calling and what their role on a specific job was. It could then be determined by the recipient if the call required an immediate quick answer, in which case it was answered with only a small pause to the task at hand, or if it was urgent and required a substantial break until the matter was resolved. What seemed to be problematic was "this automatic reaction to answer the phone" (director of a large construction company).

Instantaneity (Immediacy)

The quality of being instantaneous; instantaneousness. (OED online)

The Carnival—like the construction of a building—is a momentary occurrence, and it is always in flux, changing and evolving; as a result, the ability to make an instant connection is perhaps highly beneficial. An immediate discussion or the ability to make an instant order is strongly valued, particularly by the micro-contractor who may have three or four projects running concurrently: "I can get a call from one of the other jobs about something that has come up and I can give them an answer straight away … then the job can keep moving … do you know what I mean?" SF gives this example: "Yesterday … someone told me we were running low on nails, I forgot until I was in the car and phoned admin to order them and here (holds up an invoice) is the docket confirming the order."

The mobile phone affordance of going directly to 'the top of the pile' would seem to be complemented—in terms of problem solving at least—by the ability to immediately resolve or pass the query from the top of the pile to whomever is appropriate—like the 'game of catch' (Attali, 1985), where the 'ball' is constantly moving and changing hands, only when it stops is it forgotten about.

This immediacy however also has the tendency to act as a displacement activity and a distraction, which can of course be ameliorated by the porosity of the mobile. Opinion is divided however in its effects on forward planning. EMC, the director of a large construction company, feels there is a tendency to use the phone as a substitute for planning, leaving things to the last minute and dealing with them on the phone. As a result EMC is quite proactive in methods of training new management staff. The micro-contractor however feels they enhance his planning, he can keep update, make orders, and change his strategy during the course of the day as he gets updates from other sites.

CONCLUSION

While the contact afforded by mobile ICTs would certainly seem to enhance communication within the construction industry, it is undoubtedly in certain environs dangerous to allow devices that cause distraction. With caller id, they also present the opportunity to avoid or displace people or subjects that are particularly problematic to the receiver; unlike formal means of communication, no formal record exists of the process of attempted communication.

Apart from the health and safety issues of distraction, the next major concern with the use of mobile phones was for organization and planning. While the micro-contractor and small subcontractor felt it perhaps enhanced the particular nature of their business, the director of a large construction company felt it was in danger of being used as a substitute for forward planning; while this is discouraged, there is perhaps the anxiety that such behavior undermines the procedural aspects of day-to-day working and—more importantly—the management skill of individuals.

There is no prescription here for successful implementation of mobile ICTs, just the realization that the dispersal of the mobile phone across many groups and participants within the construction industry has affected the dynamics of communication within construction and thus affected construction itself. We suggest alternative models might give insights into understanding communication on the construction site and different technologies' effects on it. Even usage of the common phone is both diverse and complex, yet the features of the mobile phone are undoubtedly still underutilized. Our work continues to look into how we might use these devices as portals for site workers to access relevant information.

ACKNOWLEDGMENT

The themes in this article have been inspired and nurtured by discussions with Richard Coyne and John Lee. Thanks must also be extended to E. Mc Kenna, Sean Farrell, and Cyril Ronaghan for their insights into their business and practice, and for their time and patience.

REFERENCES

Attali, J. (1985). *Noise: The political economy of music* (B. Massumi, trans., vol. 16). Manchester University Press.

Bakhtin, M. (1984). *Rabelais and his world* (H. Iswolsky, trans.). Bloomington, IN: Indiana University Press.

COMIT. (2004). *Site design problem resolution.* Retrieved June 5, 2005, from http://www.comitproject.org.uk/downloads/processMaps/narratives/p7.pdf

Cox, S., & Hamilton, A. (1995). *Architect's job book* (6th ed.). London: RIBA.

Coyne, R., Lee, J., & Parker, M. (2004). Permeable portals: Designing congenial Web sites for the e-society. *Proceedings of the IADIS International Conference: E-Society,* Availa, Spain.

Deleuze, G. (1988). *Bergsonism.* New York: Zone Books.

Hyde, L. (1998). *Trickster makes this world: Mischief, myth, and art.* New York: Farrar, Straus and Giroux.

Jones, L. (2000). *The hermeneutics of sacred architecture: Experience, interpretation, comparison*. Cambridge, MA: Distributed by Harvard University Press for the Harvard University Center for the Study of World Religions.

McMeel, D., Coyne, R., & Lee, J. (2005). Talking dirty: Formal and informal communication in construction projects. *Proceedings of CAADFutures: Learning from the Past,* Vienna.

Ortony, A. (1979). *Metaphor and thought.* Cambridge: Cambridge University Press.

Peansupap, D. V., & Walker, P. D. H. T. (2005). Factors enabling information and communication technology diffusion and actual implementation in construction organizations. *ITcon, 10,* 193-218.

Pedreschi, R. (2000). *Eladio dieste. The engineer's contribution to contemporary architecture.* London: Thomas Telford.

Reddy, M. J. (1979). The conduit metaphor: A case for frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284-324). Cambridge: Cambridge University Press.

Schön, D. (1979). Generative metaphor: A perspective on problem-setting in social policy. In A. Ortony (Ed.), *Metaphor and thought.* Cambridge: Cambridge University Press.

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication* (reissue). Urbana: The University of Illinois Press.

Wiszniewski, D., Coyne, R., & Christopher, P. (1999). Turing's machines. *Proceedings of the Architectural Computing from Turing to 2000,* Liverpool.

KEY TERMS

Anthropology: Much of the theoretical underpinning for this article comes from anthropological sources, writers who have studied social and cultural phenomenon. We suggest that this pool of information has much to contribute to the discussion of mobile ICTs in construction.

Bakhtin: Writer who has extensively studied and written on Carnival in its most primitive form.

Carnival: This is taken in the Bakhtinian sense, a primal event in which everyone participates, there is no differentiation—in true Carnival—between spectator and performer, both are one in the same, as such new forms of communication would emerge.

Construction: The process of building a building, and more generally, making something.

Dirt: Anthropologically, 'dirt' is a potent metaphor which we invoke here as a representation of things which do not have allotted place. It represents material which is non-placed and problematic for systems to handle.

Information and Communication Technology (ICT): Term usually associated with e-mail, phones, skype, and so on; a technology used for communication. Not however exclusive to digital means, for example a pen could be an ICT.

Micro-Contractor: A term used here to identify a specific sort of building contractor. This type is a contractor/entrepreneurial company of 7-10 people. The size seems important to the adoption and seeming benefits of using ICT.

Mobile Knowledge Management

Zuopeng (Justin) Zhang
Eastern New Mexico University, USA

Sajjad M. Jasimuddin
University of Wales – Aberystwyth, UK

INTRODUCTION

Mobility appears to be the most important organizational and technological trend in the foreseeable future. According to IDC, two-thirds of U.S. workers will be mobile by 2006. The increasing mobility of workforces and knowledge seems to pose new challenges for organizations to effectively manage knowledge assets (cited in Mobile & Wireless Advisor, 2002). Recent advances in mobile computing make it possible for workers to use information and knowledge resources for business virtually from everywhere, which creates opportunities for organizations to engage in mobile knowledge management (mKM).

This article attempts to propose mKM strategies by studying how mobile knowledge assets can be leveraged and mKM processes can be incorporated into the main KM processes by looking at two aspects of knowledge mobility organizations. The knowledge retention processes will also be the focus to further analyze how to merge the knowledge retention processes into the main KM processes. Finally, the integration of mKM strategy into the main knowledge management processes within organizations will be explored.

The rest of the article proceeds as follows. The next section discusses the background followed by the main focus of the article, presenting strategies of mobile knowledge management and implementations. The article concludes by outlining future directions.

BACKGROUND

mKM has become an emerging business practice for knowledge management within organizations. Several studies have also touched upon the notion of mKM recently. For instance, Mummy (n.d.), an EU-based information technology company, proposes a detailed mKM model discussing related key concepts. Loutchko and Birnkraut (2005) suggest the application of mobile knowledge portals in managing mobile knowledge assets. Grimm, Tazari, and Balfanz (2002) propose a technical framework for mobile knowledge management by identifying the abstract use cases of mKM systems and by outlining a reference model that can be used to validate mKM concepts in their system architectures.

Derballa and Pousttchi (2004) examine how mobile technologies can be used to support KM and particularly to support the knowledge distribution process. With the general introduction of relevant mobile technologies, they apply the theory of mobile-added values to analyze how mobile technologies contribute to the support of KM processes. Fagrell (2000) looks at various issues related to mKM from an Informatics perspective, which includes empirical studies of mobile work, technologies for mKM systems, and the design and validation of prototype systems, observing mobile service electricians and mobile news journalists. He develops and refines mobile technologies to be used in knowledge systems, and also makes a contribution in outlining a generalized technological architecture that can be applied to mobile work settings. However, managing mobile knowledge assets are not explicitly investigated. The article will partially fill that gap by studying how mobile knowledge assets can be leveraged and how the mKM process can be incorporated into the main KM processes.

STRATEGIES OF MOBILE KNOWLEDGE MANAGEMENT

In this section, the mKM strategies will be discussed focusing on how mobile knowledge assets can be leveraged and mKM processes can be incorporated into main KM processes. To discuss mKM strategies, two special characteristics of mobile knowledge management need to be considered: the mobility of organizational knowledge and that of corporate environment. The mobility of organizational knowledge may result from the innate nature of knowledge as well as high turnarounds of workforces within organizations. As a result, an additional knowledge retention process is required to fully leverage the mobile knowledge assets, as shown in Figure 1. Environmental mobility within organizations determines the special features of mKM processes from the main KM processes. mKM processes are shown to have the similar components and flows as those of main KM processes. However, each sub-process within the mKM processes may have distinctive features. Therefore, mKM should be incorporated into the main knowledge management processes of organizations considering its special features.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. mKM: Leverage and integration

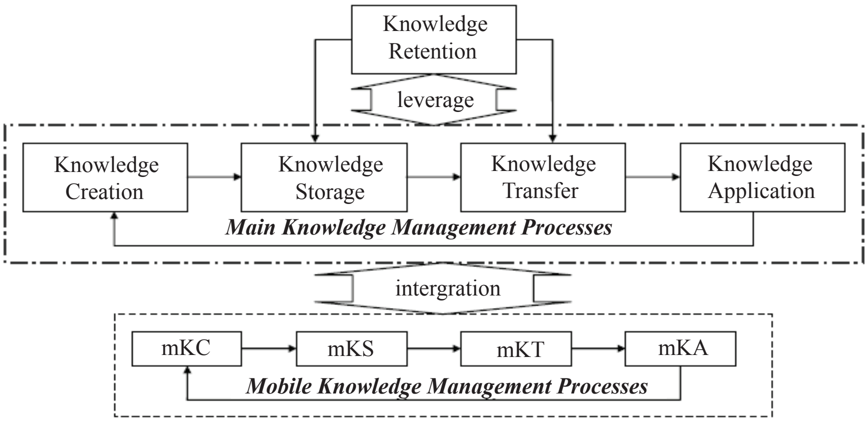


Figure 2. The relationship of m[KM], [mK]M, and KM

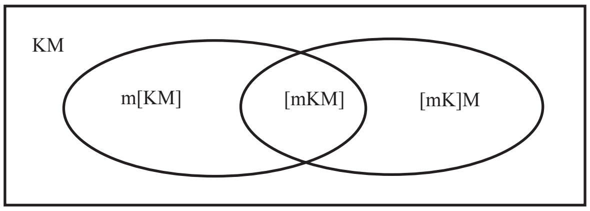


Figure 3. The relationship between [mK]M and m[KM]

		Dimensions of environmental mobility	
		Time	**Location**
Dimensions of knowledge mobility	**Detachability**	Upradability	Generality
	Volatility	Longevity	Immutability

KM, [mK]M, m[KM], and [mKM]

KM is used to denote the entire knowledge management processes within organizations. To investigate mKM strategies, the implications of mobile knowledge management will be discussed from the following three subsets:

1. **m[KM]:** mobile "Knowledge Management"
2. **[mK]M:** "mobile Knowledge" Management
3. **[mKM]:** intersection of m[KM] and [mK]M.

The first subset m[KM] focuses on the mobility of knowledge management processes due to the mobile KM environment of organizations. In particular, this subset deals with the management and integration of mobile processes of knowledge in mobile environments. The second subset [mK]M emphasizes knowledge management from the perspective of the knowledge mobility in organizations. Specifically, the second subset manages and leverages knowledge assets by taking into account the innate mobile nature of knowledge. The third subset is the intersection of m[KM] and [mK]M, which considers the mobility issues of both organizational knowledge and corporate environment.

As shown in Figure 2, some KM processes belong to m[KM] processes, some to [mK]M, and others to [mKM]—the intersection of m[KM] and [mK]M. The degree of knowledge mobility refers to the innate property of knowledge. The second refers to the increasing mobile environment of knowledge management.

UGLI: Relationship between [mK]M and m[KM]

Based on the prior discussion on [mK]M and m[KM], the relationship between m[KM] and [mK]M is investigated in this section. The mobility of knowledge is considered from the following two aspects: detachability and volatility. Detachability denotes the degree of knowledge to be detachable and applicable in the mobile environment. Volatility means that knowledge may not be captured and retained in a timely manner and in its completeness due to the innate nature of knowledge, the turnaround of organizational members, and the mobility of environments. Although these two aspects describe different characteristics of mobile knowledge, they are essentially related—knowledge with high volatility is difficult to be detached, and knowledge with high detachability must be relatively stable.

The environmental mobility is also measurable from two dimensions, time and location, which are then related with the two dimensions of knowledge mobility as shown in Figure 3. The relationship between knowledge mobility and

Figure 4. mKM strategies

		Mobility of knowledge	
		Low	High
Mobility of environment	Low	KM	[mK]M
		Conservation	*Leverage*
	High	m[KM]	[mKM]
		Integration	*Combination*

environmental mobility can be denoted by UGLI—that is, upgradeability, generality, longevity, and immutability:

- **Upgradability:** Measures knowledge detachability against the time dimension under a mobile environment. Knowledge with low upgradeability can be applied to solve problems with few upgrades over certain periods of time.
- **Generality:** Implies the connectivity of modularized knowledge with various situations. In other words, how well detached knowledge may be applied under different contexts or conditions. Knowledge with high generality can be applicable with few modifications for various contexts.
- **Longevity:** Denotes the stability of knowledge over certain periods of time as well as knowledge sensitivity of being captured against time.
- **Immutability:** Describes the volatility of knowledge against the location dimension of the mobile environment. Knowledge with higher immutability will be easier to be captured from various locations, codifiable and stored, and then transferred to other locations for applications.

Mobile knowledge with high detachability normally has low upgradeability and high generality, and mobile knowledge with high volatility normally has low longevity and immutability.

CLIC: mKM Strategies and Implementation

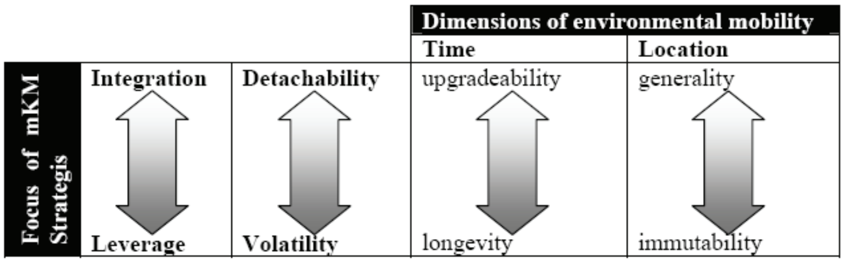
Having discussed the relationship between [mK]M and m[KM], the focus will be on the strategy of mobile knowledge management and its implementation. Based on the degrees of knowledge mobility and environmental mobility, four situations of knowledge management can be summarized with different focuses: KM, [mK]M, m[KM], and [mKM]. For each of these four cases, the knowledge management strategies are proposed as shown in Figure 4.

The strategies of mobile knowledge management can be categorized as CLIC—that is, conservation, leverage, integration, and combination:

- **Conservation:** When both the mobility of knowledge and that of environment are low, the firm should focus on its traditional knowledge management strategy.
- **Leverage:** When the mobility of knowledge is high and that of environment is low, the firm should modify its original strategy to leverage its mobile knowledge assets.
- **Integration:** When the mobility of knowledge is low and that of environment is high, the firm needs to adjust its original strategy to integrate its mobile knowledge management processes into its main processes.
- **Combination:** When both the mobility of knowledge and that of environment are high, the firm should combine both the leverage and integration strategies into its original knowledge management strategies.

Among these four strategies, integration and leverage are the bases of managing mobile knowledge assets. Although they have different focuses, both integration and leverage strategies have to deal with the knowledge mobility issue while taking into account the environmental mobility. However, the integration strategy focuses more on the detachability dimension, whereas the leverage strategy focuses on the volatility dimension. As demonstrated in Figure 5, companies implementing integration strategies should draw more attention to the issues of upgradeability and generality

Figure 5. The focus of mKM strategies



of knowledge mobility, whereas those implementing leverage strategies should concentrate on the issues of longevity and immutability.

Mobile technologies and devices enable the fulfillment of integration and leverage strategies. As described by Derballa and Pousttchi (2004), there are four major attributes of mobile technologies that add value to its original networks: ubiquity, context-sensitivity, identifying functions, and command and control functions. These special attributes of mobile technologies support mKM strategies to increase knowledge detachability and lower knowledge volatility. The interfacing processes of mKM strategies between mobile KM and main KM processes may include uploading, downloading, and synchronization, which connects mobile users with the main knowledge base so that mobile knowledge management processes can be integrated and mobile knowledge can be leveraged.

Considering a firm that implements mKM strategies through mobile technologies, it will have to make the following decisions:

1. how much to invest on mobile technologies to implement mKM strategies, and
2. how to balance the tradeoffs between integration and leverage strategies.

The first decision may be made by analyzing the total impact of mKM strategies on the firm's business goal and its original KM plans. The second decision can be made based on both the degree of knowledge mobility and environment mobility, and the actual effects that may be achieved from integration and leverage strategies. As noted earlier, companies emphasizing knowledge mobility should focus on leverage strategy, whereas those emphasizing environmental mobility should focus on integration strategy. In addition, the effects of technology investments on both strategies have to be considered. For instance, it is relatively more difficult to reduce knowledge volatility by investing in mobile technologies because the innate nature of knowledge mobility is immutable. Finally, the transition between two strategies should also be taken into account. Knowledge with low volatility may be easy to be detached. Hence, the implementation of leverage strategy may help fulfill integration strategy by increasing knowledge detachability.

FUTURE TRENDS

The proposed mKM strategies for organizational knowledge management are by no means the best strategies available to manage mobile knowledge assets with the help of mobile technologies. However, such mKM strategies provide great insights for better understanding of the relationship between mKM and knowledge management, which also lays the solid foundation for related future research.

First, more discussion can be made on the relationship between the detachability and volatility dimensions for the issue of knowledge mobility. The transition between these two dimensions is also worthwhile for further investigation, which may clarify the effects achieved through the enablement of mobile technologies.

Second, the dimensions used to measure the mobility of organizational knowledge may be compared with other dimensions of knowledge in organizations, such as tacitness (tacit or explicit) and location (exogenous and endogenous), as espoused by Jasimuddin (2005), which will provide a better understanding of the nature of mobile knowledge assets.

Third, related analytical models can be formulated along with the mKM strategies and decisions discussed in this article. Taking into account the relationships and effects of m[KM] and [mK]M, an analytical model may provide insights for better understanding of the intricacies of various factors influencing the management of mobile knowledge assets.

Finally, the strategic model of mKM may be empirically tested to further reveal the critical factors of the implementations of mKM strategies. Empirical tests may not only corroborate the usefulness of the strategies outlined here, but also help us better comprehend other related factors during the implementation of the strategies.

CONCLUSION

Knowledge management has been regarded as an important business practice for organizations to gain competitive edge. With corporate environments becoming increasingly mobile, organizations are seeking innovative ways to manage mobile knowledge assets. Four different situations of knowledge management—KM, m[KM], [mK]M, and [mKM]—within organizations are categorized by taking into account both the mobility of organizational knowledge and that of the corporate environment, which lays the foundation for our further discussion on the strategies of mobile knowledge management. Most specifically, the UGLI relationship between m[KM] and [mK]M is proposed based on two dimensions of knowledge mobility and environmental mobility. The UGLI (upgradeability, generality, longevity, and immutability) relationship suggests how mobile knowledge assets are perceived and interpreted under mobile corporate environments. Moreover, the CLIC strategies are outlined for mobile knowledge management according to the four situations of knowledge management. The CLIC (conservation, leverage, integration, and combination) strategies are based on how to manage mobile knowledge under mobile environments. Such a discussion of strategies for mobile knowledge management seems to provide valuable insights and guidelines for managers to adopt and implement appropriate organizational strategies by focusing on various aspects of mobile knowledge assets.

REFERENCES

Derballa, V., & Pousttchi, K. (2004). Extending knowledge management to mobile workplaces. In M. Janssen, H. G. Sol, & R. W. Wagenaar (Eds.), *Proceedings of the 6th International Conference on Electronic Commerce (ICEC'04)*.

Fagrell, H. (2000). *Mobile knowledge*. Doctoral dissertation, Department of Informatics, Göteborg University, Sweden.

Grimm, M., Tazari, M. R., & Balfanz, D. (2002, December). Towards a framework for mobile knowledge management. *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management*, Vienna, Austria. Berlin: Springer-Verlag.

Jasimuddin, S. M. (2005). External sources of organizational knowledge. In M.A. Rahim & R. T. Golembiewski (Eds.), *Current topics in management* (Vol. 10, pp. 13-37). New Brunswick, NJ: Transaction Publishers.

Loutchko, I., & Birnkraut, F. (2005). Mobile knowledge portals: Description schema and development trends. *Proceedings of the 2005 International Conference on Knowledge Management (I-Know'05)*.

Mobile & Wireless Advisor. (2002). *105 million mobile U.S. workers by 2006*. Retrieved from http://mobileadvisor.com/Articles.nsf/aid/HARTP144

Mummy. (n.d.). *Mobile knowledge management—Parts 1&2* (white paper). Retrieved from http://mummy.intranet.gr/publications.html

KEY TERMS

Detachability: Denotes the degree of knowledge to be detachable and applicable in the mobile environment.

Knowledge Management: The process of capturing organizational members' knowledge and then interpreting, transmitting, preserving, and refining it for organizational use.

Knowledge Mobility: The movement of organizational knowledge that may result from the innate nature of knowledge as well as high turnarounds of workforces within organizations.

Knowledge Retention: Referred to as the preservation of knowledge by an organization either in its people's heads, in databases, or on paper.

Mobile Knowledge Management: The process of managing mobile knowledge assets in mobile environments due to increasing mobility of the workforce.

Organizational Knowledge: Information interpreted by organizational members that helps to make effective action for the organization.

Volatility: Means that knowledge may not be captured and retained in a timely manner and in its completeness due to the innate nature of knowledge, the turnaround of employees in organizations, and the mobility of environments.

Mobile Learning

David Parsons
Massey University, New Zealand

INTRODUCTION

Mobile learning (variously shortened to M-Learning, M-learning, m-learning, mlearning, M Learning, or mLearning!) describes any form of education or training that is delivered using some kind of mobile device. As the power and sophistication of mobile devices increases, and wireless networks become faster and more ubiquitous, learning with a mobile device will become an integral part of the general spectrum of technology-supported learning. Furthermore, the special characteristics of mobile learning, including ubiquity, convenience, localization, and personalization, give it unique qualities that help it stand out from other forms of learning.

For some, mobile learning is simply an extension of electronic learning (e-learning), indeed it is sometimes referred to as mobile e-learning, and both can be seen, conceptually at least, as subsets of distance learning (Georgiev, Georgieva, & Smrikarov, 2004). However, such an approach fails to take into account either the restrictions of the mobile device, the special circumstances of the mobile learner, or the value-added aspects of mobility such as on-demand learning, ad-hoc networking, and location and context awareness.

Unlike desktop electronic learning environments, mobile learning can take place in changing contexts and via a range of mobile devices with many variations in form factor. Mobile learning systems must be adaptive both to the learner (in terms of his or her developing learning model and profile) and to the device (in terms of its functionality and its current environment). Therefore the hardware and software architectures for a mobile learning system provide a major technical challenge. Overcoming these difficulties can, however, enhance the learning experience, since mobile learning has the benefits of mobility and its supporting platform, which can be summarized as being ubiquity, convenience, localization, and personalization. Ubiquity means that the learning content can be accessed anywhere, regardless of location. With ever-increasing coverage by mobile network providers, mobile learning services can have an increasingly ubiquitous presence. This is particularly useful where mobile learning provides support for learners in the field, where information is urgently needed on site, perhaps to help diagnose medical conditions, analyze field data, or repair equipment. As mobile devices become more pervasive in our everyday lives, using these devices for learning becomes more convenient. High-speed 'always-on' data connections mean that the learner can access material when it is convenient to them, enabling learning to take place on an ad hoc basis.

Localization is a specific strength of mobile devices, since they can use location awareness to provide services that are targeted to the user's current locality. Location awareness has been used in a number of mobile learning solutions to enhance the user's experience, for example in the Ambient Wood project, where children explored a woodland environment using mobile technology (Rogers et al., 2004). Finally, personalization is a key component of mobile learning for two reasons. First, the difficulty of navigation and small screen size of mobile devices means that it is important to target learning material to the user as much as possible. Second, such targeting is easier for enrollment-based services like education, where the provider is likely to be able to gather considerable information about learners and construct accurate profiles of their activities and requirements.

In defining mobile learning we should also be aware that mobile learning does not necessarily mean wireless learning. A number of mobile learning solutions, typically those deployed on personal digital assistants (PDAs), enable the user to install a complete learning module that runs as a standalone application on the device. This type of approach is used in mobile applications that act as simple electronic training manuals, as well as tourist and museum guides, where the requirement for the software is short term and context driven, though location awareness may not be supported in non-wireless systems. While these types of mobile learning applications can be valuable, systems that include wireless connectivity can provide more interactivity, context awareness, and learner choice.

USING THE MOBILE ENVIRONMENT FOR LEARNING

As well as solving technical problems, developing a successful mobile learning solution depends on imaginative use of the mobile environment. There are examples in the literature of systems that leverage mobility to mimic film narrative, integrate mobile devices into multi-technology environments, encompass group game play, and store material in virtual spaces. In many cases these systems take advantage of location awareness and the ability of wireless devices to support communication between group members. Thus mobility enables individuals to participate in distributed simulations

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

and role plays across both space and time. Examples include the use of mobile devices to model the spread of disease and enable multi-role simulations. Learning scenarios with multiple branches can be used to indicate the usefulness of different outcomes, depending upon the decisions made by the learner, for example in medical diagnosis (Setaro, 2001). Lundin and Nulden (2003) describe multimedia scenarios used in a professional context based on the PIER approach, which has four main building blocks: problem-based learning, interactive multimedia, experiential learning, and role playing.

A number of examples of mobile learning are driven very much by the context of the learner, where the mobile device can be taken into an environment and be used as learning support in that environment. Examples of this type of system include those that provide location-aware information that is relevant to the context, those that allow the sharing of information about related contexts, and those that provide information about a context, regardless of actual location. *Location-aware* systems that provide information include tourist systems, and museum and archaeological systems. An interesting example is Urban Tapestries, which links urban stories to specific London locations (Walker, 2004). An early but interesting example of sharing information about related contexts is Wireless Coyote, where mobile learners working in different areas assisted each other's analyses of the environment (Grant, 1993). Sharples (2002) refers to this type of mobile learning system as conversational learning, stressing the value of interactivity between mobile learners. Systems that provide context-related information regardless of actual location include a mobile bird-watching learning system (Chen, Kao, Sheu, & Chiang, 2002) and ELDIT, a language translation system (Trifonova, Knapp, Ronchetti, & Gamper, 2004).

TYPES OF MOBILE LEARNING DELIVERY

Mobile learning can support many types of learners who will require different types of content and modes of delivery, and a mobile system may integrate with a wider academic or professional program or it may be stand alone. To support the flexibility of the learner, the content of a mobile curriculum needs to be broken down into short and focused nuggets of learning, the type of content that can be accessed in the 'downtime' of the learner, particularly in mobile learning systems that target the professional in the field. In situations where learning is used in a professional context, there is a concept of *just-in-time learning* (Koschembahr, 2005), where the mobile learning content has an on-the-job training focus. In this vein, workers in industries such as retail and fast food can get what might be called *fast learning* (McGee, 2003), focusing on low-level training modules and product information. A characteristic of this approach is the blurring of boundaries between acquiring information and learning. This encourages the view that, before long, an employee will not even be able to differentiate learning from other everyday job functions. However, we might question whether such a system is actually delivering learning, since the information being conveyed is often transient and trivial.

MOBILE LEARNING AS A DISRUPTIVE TECHNOLOGY

One of the key characteristics of a disruptive technology is that its adopters are prepared to accept a reduction in some qualities in order to benefit from innovation (Funk, 2004). There are some positive indications that mobile learning can be successful even in limited technical environments, suggesting that it has the characteristics of a disruptive technology. Research has shown that learners are prepared to accept technological limitations for the benefits of mobility. A study by Ericsson in 2002 showed that even with a simple wireless access protocol (WAP) browser interface, users felt that mobile learning could be a quality experience (Ericsson, 2002). In the study, 77% of participants felt that mobility actually increased the quality of electronic learning, and all felt that one of its key benefits was its ability to increase access to education and training. Another study using the short message service (SMS) as an interactivity mechanism also suggested that the lack of sophistication of the platform need not be a major stumbling block to the quality of the learning experience. This study indicated that mobile learning applications can have depth and complexity, and encourage wider scale participation, even where it might be expected that technical limitations would discourage the learner (Stone, Briggs, & Smith, 2002). It seems therefore that technological sophistication is not necessarily a measure of usefulness, since even simple technologies like classroom response systems have proved effective, engendering rich social practice around basic systems (Roschelle, 2003).

CONCLUSION

In summary, mobile learning is a specialized field from both a technical and educational perspective, and one that will become increasingly important as wireless communication networks and mobile devices become more pervasive and sophisticated. In many ways, electronic and mobile learning will move closer together as the power and sophistication of mobile devices increase. However there will always be certain aspects of mobility, in particular ubiquity and location awareness, that will make mobile learning a unique and special approach to education.

REFERENCES

Chen, Y., Kao, T., Sheu, J., & Chiang, C. (2002). A mobile scaffolding-aid-based bird-watching learning system. *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'02)* (p. 15). Växjö, Sweden.

Ericsson. (2002). *Mobile learning in action: Report on the use of mobile telephones for training*. Retrieved January 24, 2006, from http://learning.ericsson.net/mlearning2/project_one/mobile_learning.html

Funk, J. (2004). *Mobile disruption.* Hoboken, NJ: John Wiley & Sons.

Georgiev, T., Georgieva, E., & Smrikarov, A. (2004, June 17-18). M-learning: A new stage of e-learning. *Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech2004)* (pp. 1-5). Rousse, Bulgaria.

Grant, W. C. (1993). Wireless Coyote: A computer-supported field trip. *Communications of the ACM, 36*(5), 57-59.

Koschembahr, C. (2005). *Optimizing your sales workforce through mobile learning.* Retrieved September 2005 from http://www.learningcircuits.org/2005/apr2005/vonKoschembahr.htm

Lundin, J., & Nulden, U. (2003). Mobile scenarios: Supporting collaborative learning among mobile workers. In *Educating managers with tomorrow's technologies* (pp. 173-190). Greenwich, CT: Information Age Press.

McGee, M. K. (2003). *E-learning on the fly.* Retrieved September 2005 from http://www.informationweek.com/story/showArticle.jhtml?articleID=15800505

Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith, H., et al. (2004, June 1-3). Ambient Wood: Designing new forms of digital augmentation for learning outdoors. *Proceedings of the 3rd International Conference for Interaction Design and Children (IDC 2004)* (pp. 1-9). College Park, MD.

Roschelle, J. (2003). Unlocking the learning value of wireless mobile devices. *Journal of Computer Assisted Learning,* 19, 260-272.

Setaro, J. L. (2001). *If you build it, will they come? Distance-learning through wireless devices.* Retrieved September 2005 from http://www.unisysworld.com/monthly/2001/07/wireless.shtml

Sharples, M. (2002). Disruptive devices: Mobile technology for conversational learning. *International Journal of Continuing Engineering Education and Lifelong Learning, 12*(5/6), 504-520.

Stone, A., Briggs, J., & Smith, C. (2002). SMS and interactivity—Some results from the field, and its implications on effective uses of mobile technologies in education. *Proceedings of the IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'02)* (p. 147). Växjö, Sweden.

Trifonova, A., Knapp, J., Ronchetti, M., & Gamper, J. (2004). Mobile ELDIT: Transition from an e-learning to an m-learning system. *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 188-193).

Walker, K. (2004). Learning on location with cinematic narratives. *Proceedings of the 1st Workshop on Story Representation, Mechanism and Context* (pp. 55-58). New York.

KEY TERMS

Conversational Learning: Using mobile devices as an aid to learning by interacting with other learners.

Localization: The delivery of services to the user that are aware of the user's current location and therefore tailored to that context.

Mobile Learning: Learning that takes place using some kind of mobile device.

Personalization: Providing content to the user that is based on his or her user profile.

Short Message Service (SMS): A service that allows short text messages to be passed between mobile phones via unused time on the control channel, using a store and forward mechanism.

Ubiquity: The availability of a service in most, if not all locations.

Wireless Access Protocol (WAP): A communications protocol developed specifically for mobile phones which supports page markup using the Wireless Markup Language (WML).

Mobile Learning Environments

Paul Crowther
Sheffield Hallam University, UK

Martin Beer
Sheffield Hallam University, UK

INTRODUCTION

Mobile learning requires a methodology for creating mobile learning scenarios and producing learning objects to implement them. It also requires a technology to deliver the learning objects to users via mobile computing devices such as personal digital assistants, smart phones, and tablet computers. This article will describe both the pedagogic methodology and the technology using the European research project *MOBIlearn* as an example.

A key part of the MOBIlearn project is the integration of new technologies in education. It aims at improving access to knowledge for selected target users, giving them ubiquitous access to appropriate learning objects (Taylor, 2003). "The MOBIlearn project intends to develop software that supports the use of mobile devices (smart phones, PDAs, Tablet PCs and laptops with wireless network connection) for various learning scenarios, including non-institutional learning" (MOBIlearn, 2005). The aim of MOBIlearn is therefore "… the creation of a virtual network for the diffusion of knowledge and learning via a mobile environment … to … demonstrate the convergence and merging of learning supported by new technology, knowledge management, and new forms of mobile communication" (MOBILearn, 2002, Annex 1, p. 7).

The pedagogic aim of the system is to provide users with the ability to engage in formal, non-formal, and informal learning in a personal collaborative virtual learning environment. To this end, three scenarios were used as the basis of developing the requirements for the system. These were a formal university course and a related orientation activity, a non-formal health care scenario, and an informal scenario based around museums and galleries. The requirements of the system were captured from three user-developed scenarios. A use case model was produced for each of these scenarios plus a fourth model describing generic or common requirements. These requirements were further documented for the technical developers in a database based on a Volere template (Robertson & Robertson, 2001).

The philosophy behind the MOBIlearn system is that it provides a set of interoperable services. Services should be able to communicate asynchronously using unstable communication channels (MOBIlearn, 2005). At the center of the system is the component providing the portal services including the main portal. This represents the single access point for the user to all the services provided by the MOBIlearn system.

One of the challenges of creating a mobile learning environment that spanned more than one domain was extracting generic requirements applicable to all domains. A corollary to this was identifying requirements that were specific to a domain.

Scenarios

Initially the MOBIlearn requirements were provided by three scenarios:

1. A visit to an art gallery.
2. Access to training and basic medical knowledge in a hospital.
3. Master's in business administration.

Development of these scenarios was essential for deriving both user and technical requirements.

In the MOBIlearn project a series of distributed development teams were established with specific roles or workpackages. One of these workpackages involved the development of requirement specifications to be used by the technical workpackage teams.

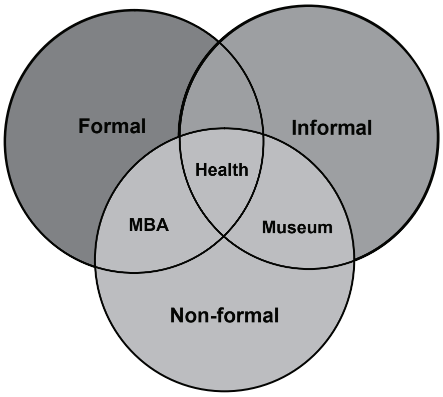
Initially these requirements were derived from the user scenarios listed above using the use cases alone, one set for each scenario. The next stage was to amalgamate these into a single specification which could then be handed on to the software developers for the final system. The technical teams then developed a series of services which were required to implement the use cases.

MOBIlearn Pedagogic Design

The pedagogic basis of the system is the learner who interacts with the mobile learning portal to access learning objects and participate in online activities. MOBIlearn provides a tool to facilitate collaboration and teamwork. It expands on systems such as OTIS (Occupational Therapy Internet School) (Beer, Slack, & Armitt, 2005) to provide a framework that can be

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Types of learning and their relationship to the scenarios



used in variety of learning situations. It also allows a variety of learning styles.

Learners today want to learn when and where they want, in formal, non-formal, and informal ways (Brand, Petrak, & Zitterbart, 2002; Cook & Smith, 2004).

The types of learning are characterized by the following attributes:

- **Formal**
  - Mandatory participation
  - Objectives and means controlled by educator
- **Non-Formal**
  - Voluntary participation
  - Objectives controlled by learners
  - Means controlled by educator
- **Informal**
  - Grows out of spontaneous situations
  - Objectives and means controlled by learners
  - There may be a facilitator who may provide some content

The second feature of the environment is that it facilitates communities of learners. In the case of the museum scenarios, the learners are operating in an informal environment motivated by their own interests (Cook & Smith, 2004). The methodology gives them the ability to join a virtual community with interests like their own. The learner is under no obligation to formally join (or leave) the community and can participate as much or as little as he or she wishes. This particular scenario has many features in common with the Virtual Museum of Canada (Soren, 2005), but is also designed to be used in a real museum (the Uffizi Gallery in Florence, Italy, being a test site) to give a richer experience than the traditional audio guides.

The health care scenario on the other hand is a non-formal learning environment where a *community of practice* (CoP) is being established. The system is designed to deliver training case studies where a learner assesses a situation and suggests a course of action. This can then be discussed with other learners who will have different levels of experience. Learning has no start or end point, and new members can join (and leave) at any time; however, it may be a condition of employment that staff engage with this continuing development. New case studies can be added, including ones suggested by the learners. This does contradict some of Ellis, Oldridge, and Vasconcelos's (2003) criteria for a community of practice, specifically a voluntary and emergent group. However, if staff engages with the learning environment, a virtual community of practice could develop, meeting other criteria including a mutual source of gain.

Finally there is the MBA scenario, which is based in formal learning where students use the system to access resources, undertake tasks, and discuss topics with fellow students and academics. There is immersion and presence in the online learning environment. This encourages students to build trust and teamwork (Beer et al., 2005). The environment is more constrained and there is a specific enrolment and end point. Although it is theoretically possible to start and end a course at any time, this does not yet happen.

There are features that are common to all three scenarios, for example, there will be some base content. In the case of the museums, this will be information about exhibitions and within that, information about specific exhibits. In the case of health care, there are a series of reference *oblettes* (learning objects) relating to various diseases and situations. For the MBA, there are formal course materials. All scenarios have discussion areas or forums allowing collaborative learning and providing for the foundations for a community of learning and practice to be built.

Development

The development methodology for MOBIlearn was a combination of the *service-oriented approach* (SOA) and prototyping loosely based on Boehm et al.'s (1998) spiral model. Modeling was done using the use-case modeling tool of UML (Rumbaugh, Jacobson, & Booch, 1999). Initially a use case model was produced for each of the scenarios. These were examined for commonality, and a common or generic use case diagram was produced mapping user requirements used in all scenarios (see Figure 2). For each of the use cases, a detailed description was produced based on the template suggested by Cockburn (1998) (see Figure 3).

The use case diagrams were then cross-referenced in the requirements database, which was based on Volere shells (Robertson & Robertson, 2001) (see Figure 4). The database was then passed on to the development teams who grouped requirements according to their functional similarity. This became the basis of the service-oriented architecture described as follows.

Figure 2. Top-level generic or general use case diagram for MOBIlearn

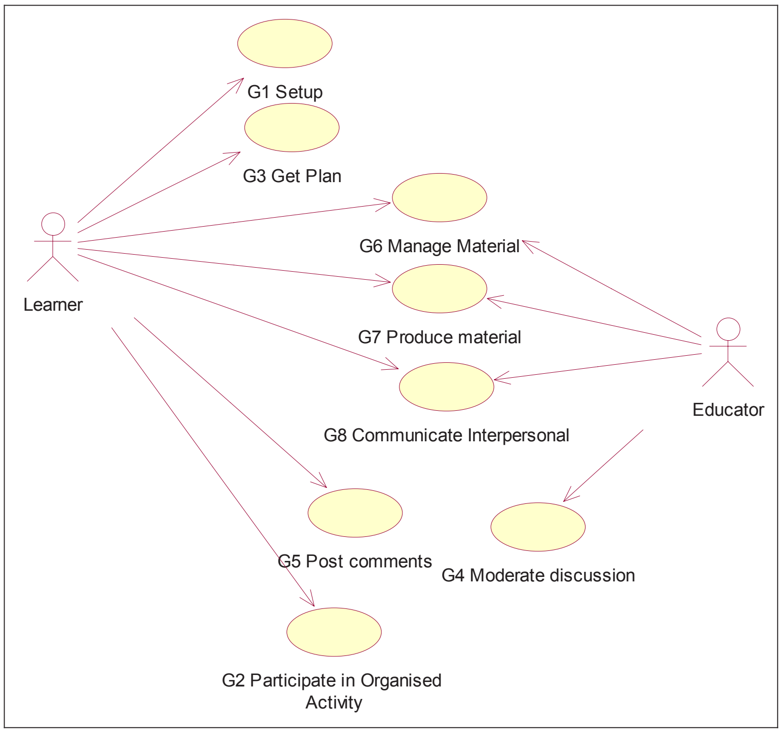
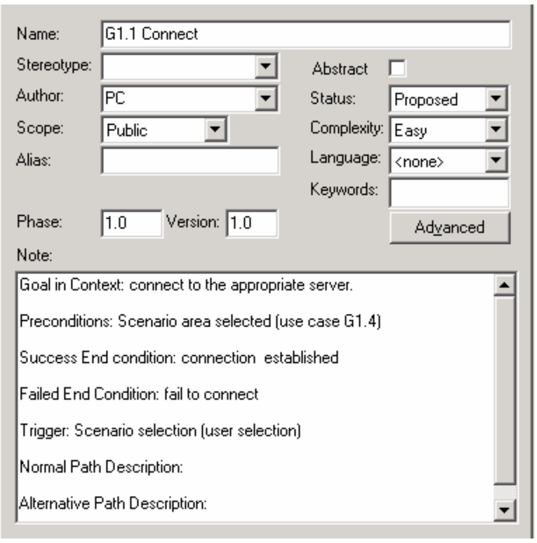

G1 Setup
G3 Get Plan
G6 Manage Material
Learner
G7 Produce material
Educator
G8 Communicate Interpersonal
G5 Post comments
G4 Moderate discussion
G2 Participate in Organised Activity

Figure 3. Use case description template


Name: G1.1 Connect
Stereotype:
Abstract ☐
Author: PC
Status: Proposed
Scope: Public
Complexity: Easy
Alias:
Language: <none>
Keywords:
Phase: 1.0
Version: 1.0

Note:

Goal in Context: connect to the appropriate server.

Preconditions: Scenario area selected (use case G1.4)

Success End condition: connection established

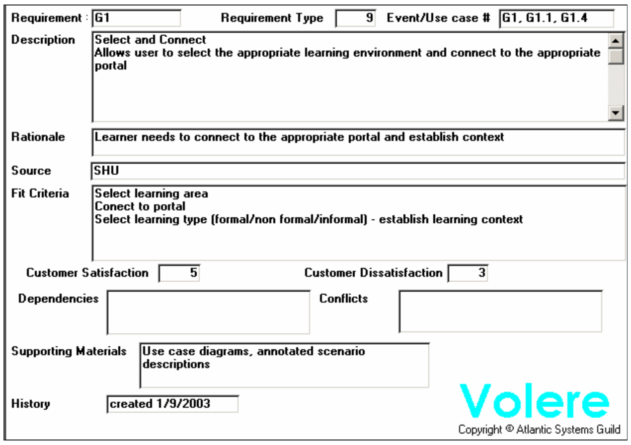
Failed End Condition: fail to connect

Trigger: Scenario selection (user selection)

Normal Path Description:

Alternative Path Description:

Figure 4. Volere template related to the use case in Figure 3

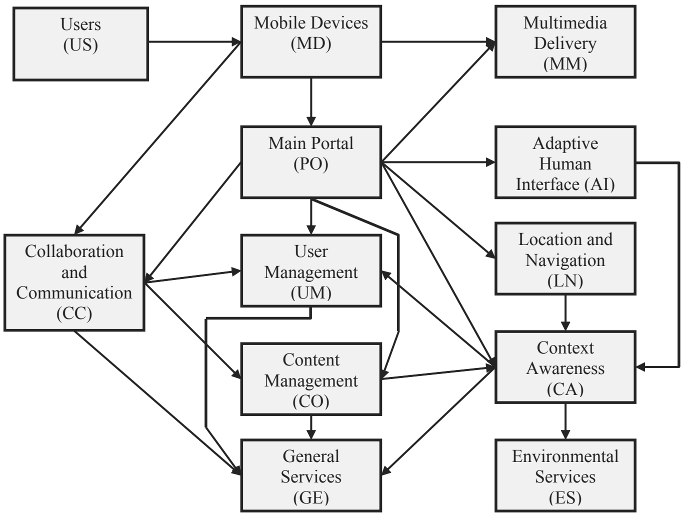

Requirement : G1
Requirement Type 9
Event/Use case # G1, G1.1, G1.4
Description
Select and Connect
Allows user to select the appropriate learning environment and connect to the appropriate portal
Rationale
Learner needs to connect to the appropriate portal and establish context
Source
SHU
Fit Criteria
Select learning area
Conect to portal
Select learning type (formal/non formal/informal) - establish learning context
Customer Satisfaction 5
Customer Dissatisfaction 3
Dependencies
Conflicts
Supporting Materials
Use case diagrams, annotated scenario descriptions
History
created 1/9/2003
Volere
Copyright © Atlantic Systems Guild

MOBIlearn Architecture

MOBIlearn is an example of a personal virtual environment (PVLE) (Xu, Wang, & Wang, 2005) consisting of domain-level knowledge from the content provider (e.g., a museum or university) and a meta-level model to allow a learner's profile to be matched to the environment and the mobile device he or she is using.

Each of the development scenarios has its own objects, but what they all have in common is that they need to be delivered in a flexible way to a variety of devices. For example the interface characteristics of a tablet computer are far different from those of a PDA. One challenge is therefore to deliver the correct interface to a learning object or oblette to the mobile device.

Figure 5. High-level component diagram in the MOBIlearn architecture (MOBILlearn documentation V 2.47, p. 32)



There are a variety of ways of delivering learning materials to devices with differing characteristics including re-authoring, transcoding, and the functional-based object model (Kinshuk & Goh, 2003). Ideally, an open standard should be used to allow different content providers to make their material available on mobile devices. The approach taken in MOBIlearn is to use re-authoring where page descriptions are held as XML which is compatible with the standard suggested by Loidl (2005).

Figure 5 shows the overall architecture of the MOBIlearn system. Users are users of the system who interact with it using a variety of mobile devices. In other words these two system components are not part of what was developed. Using their mobile devices, a user's main interaction is with the main portal service that controls access. The main portal service then calls other services depending on the user's selection. When a user logs on, the services in the adaptive human interface component are used to deliver the appropriate interface to the user's device. After logging on, a user may wish to download a learning object which will require a call to the content management component. Based on what that component sees, it may wish to engage in a collaborative activity requiring a call to the services of the communication and collaboration component.

CONCLUSION AND FUTURE DIRECTIONS

MOBIlearn provides an architecture and methodology to support users using a variety of learning styles and requirements. It can be used in a variety of learning situations ranging from formal university courses to informal communities with a common interest. Learners have an online presence and can engage in collaboration and teamwork. The services a learner needs to do this are delivered via a portal and adapted to the physical device he or she is using.

The implications of this are enormous. In the informal learning environment of a museum, a learner will no longer be dependent on audio guides and labels. The system can deliver a much richer experience, plus the learner will be able to collaborate with other learners with similar interests. Health care and other professionals will be able to use the system as a support and training tool while on the job. In the formal learning environment, the system will add more flexibility for both distance and face-to-face learners.

ACKNOWLEDGMENTS

We acknowledge the EU for financial support through the MOBIlearn project (IST-2001-37440). The views expressed in this article are those of the authors and may not represent the views of the EU.

REFERENCES

Beer, M., Slack, F., & Armitt, G. (2005). Collaboration and teamwork: Immersion and presence in an online learning environment. *Information Systems Frontiers, 7*(1), 27-35.

Brand, O., Petrak, L., & Zitterbart, M. (2002). *Support for mobile learners in distributed space.* Retrieved November

7, 2003, from www.learninglab.de/~brand/Publications/elearn02.pdf

Boehm, B., Egyed, A., Kwan, J., Port, D., Shah, A., & Madachy, R. (1998). Using the WinWin spiral model: A case study. *IEEE Computer, 31*(7), 33-44.

Cockburn, A. (1998). *Basic use case template.* Retrieved September 26, 2003, from http://members.aol.com/acokcburn

Cook, J., & Smith, M. (2004). Beyond formal learning: Informal community e-learning. *Computers and Education, 43,* 35-37.

Ellis, D., Oldridge, R., & Vasconcelos, A. (2203). Community and virtual community. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (vol. 37, pp. 145-146).

Kinshuk & Goh, T. (2003). Mobile adaptation with multiple representation approach as educational pedagogy. *Proceedings of Wirtschaftsinformatik 2003—Medien—Markte—Mobilitat* (pp. 747-763), Heidelberg, Germany.

Loidl, S. (in press). Towards pervasive learning: WeLearn. Mobile. A CPS package viewer for handhelds. *Journal of Network and Computer Applications.*

MOBIlearn. (2002). *Next generation paradigms and interfaces for technology supported learning in a mobile environment exploring the potential of ambient intelligence.* Annex 1, Information Society Technologies Program EU Proposal/Contract: IST-2001-37187.

MOBIlearn. (2005). *The MOBIlearn software documentation V 2.47.* Retrieved September 15, 2005, from http://bscw.uni-koblenz.de/bscw/bscw.cgi

Robertson, J., & Robertson, S. (2001). *Volere: Requirements specification template* (8th ed.). Atlantic Systems Guild.

Rumbaugh, J., Jacobson, I., & Booch, G. (1999). *The Unified Modeling Language reference manual.* Boston: Addison-Wesley.

Soren, B. J.(2005). Best practices in creating quality online experiences for museum users. *Museum Management and Curatorship, 20,* 131-148.

Taylor, J. (2003, May). A task-centred approach to evaluating a mobile learning environment for pedagogical soundness. *Proceedings of MLEARN 2003, Learning with Mobile Devices,* London.

Xu, D., Wang, H., & Wang, M. (2005). A conceptual model of personalised virtual learning environments. *Expert Systems with Applications, 29,* 525-534.

KEY TERMS

Community of Practice (CoP): A flexible group informally bound by common interests.

Formal Learning: Learning in a structured and controlled environment with fixed, specified learning objectives.

Informal Learning: Learning motivated by personal interest with no specific learning objective and structured by the individual or by an independent informal group.

MOBIlearn: A system that provides both a methodology and a technology to deliver flexible learning in a mobile environment.

Non-Formal Learning: Learning in a formal environment, but with no formal learning objectives.

Oblette: A learning object or self-contained piece of material designed to meet a learning objective.

Pedagogy: The set of activities of educating or instructing or teaching; activities that impart knowledge or skill.

Service Oriented: A set of interoperable services, developed independently, which interact to provide the learning environment.

Mobile Medical Image Viewing Using 3G Wireless Network

Carrison K. S. Tong
Pamela Youde Nethersole Eastern Hospital, Hong Kong

Eric T. T. Wong
Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Teleradiology is a routine practice for radiologists to make urgent diagnosis by remote viewing radiological images such as computed tomographic (CT), magnetic resonance (MR), computed radiographic (CR), and digital radiographic (DR) images outside their hospitals. Traditionally, due to limited network bandwidth and huge image file sizes, this technique was limited to fixed-point communication using an integrated services digital network (ISDN) and broadband network. Without any prior information, most radiologists would invariably require high-quality display units and lossless compressed images for their clinical diagnosis. Besides the technical issues involved in the uninterrupted provision of a 24-hour teleradiology service, most hospital administrators have to consider a series of management issues on the quality of this service such as data confidentiality, integrity, and accessibility.

This article presents the implementation process of a high-quality teleradiology service using the third-generation (3G) wireless network. In the provision of this service, several high-quality notebook computers with a 15-inch liquid crystal display (LCD) screen of resolution 1,024 x 768 pixels and 32-bit color have been configured to view medical images in the digital imaging and communications in medicine (DICOM) format using a Web browser. These notebook computers are connected with 3G mobile phones so that users could access the Internet using Web browsers through the 3G network at a speed of at least 384 kbps. The users could also use the Web browser for logging into the hospital network through an application tunneling technique in a virtual private network (VPN). When logging into the VPN, for security purposes the network authentication is enhanced by a one-time and two-factor authentication (OTTFA) mechanism. In OTTFA, the user password contains two parts: a personal password and a randomly generated password. After successfully logging into the hospital network, the user has to log into the image server using another account name and password. The above are all important to ensure the high standard of confidentiality of the system.

The data volume of the image server is about 1 TB, stored in a level-5 configured redundant array of inexpensive disks (RAID). For management of this huge amount of data, the location of each image in the storage unit is stored in a Structural Query Language (SQL)-based database. Each image also has DICOM tags for storage of the patient name, identity (ID) number, study date, and time. After the success of each login, the user can query the image server for related images using the patient's demographic data such as the study date. These are used to enhance the integrity of the system.

There are three image servers configured in a high available (HA) cluster using a load-balancing switch. The user could access any one of the servers for diagnostic purpose using the teleradiology technique. This setting is used to ensure the availability of the service 24 hours a day/7 days a week. The above system has operated for six months, and zero downtime was recorded. This leads to the belief that it is feasible to operate a quality teleradiology system using 3G networking technology with the important concerns of data confidentiality, integrity, and accessibility being dealt with in an effective manner.

BACKGROUND

Teleradiology is the process of sending radiologic images from one point to another through digital, computer-assisted transmission, typically over standard telephone lines, a wide-area network (WAN), or over a local area network (LAN). Through teleradiology, images can be sent to another part of the hospital or around the world.

In a hospital environment, it is not unusual that sometimes certain senior or experienced clinical staff would not be available onsite. These senior clinical staff may standby at home, on business trip, or just on their way to work. For urgent medical cases, remote consultation is required. It is important to have multimedia communication, including voice, text, and picture, between the senior clinicians and the hospital. A reliable, secure, easy-access, manageable, high-speed, standardized, multimedia medical consultation system is required.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

PROBLEM

Today, teleradiology is still facing many limitations such as low network bandwidth, limited locations, and implementation issues associated with security, standards, and data management.

Limited Locations and Low Network Bandwidth

Depending on data-transfer rate requirements and economic considerations, images can be transmitted by means of common telephone lines using twisted pairs of copper wire, digital phone lines such as ISDN, coaxial cable, fiber-optic cable, microwave, satellite, and frame relay or T1 telecommunication links.

Today most teleradiology systems run over standard telephone lines. Over the next couple of years, we should see a substantial migration to switched-56 and ISDN lines, which offer higher speed and better line quality than standard dial-up phone lines. Other high-speed lines, including T1 and SMDS (shared multimegabit data services), will also become more popular as prices continue to drop.

However, remote consultation on fixed lines can only be performed in pre-installed locations such as a radiologist's home. A wide-area wireless network can provide a more flexible teleradiology service for the users (Oguchi, Murase, Kaneko, Takizawa, & Kadoya, 2001; Reponen et al., 2000; Tong, Chan, & Wong, 2003).

Security in a 3G Network

The fragile security of 2.5G and 3G wireless applications was abundantly evident in Japan recently when malicious e-mails to wireless handsets unleashed a malevolent piece of code which took control of the communications device and, in some cases, repeatedly called Japan's national emergency number. Other cell phones merely placed several long-distance calls without the user's knowledge, while others froze up, making it impossible for subscribers to use any of the carrier's services. Incidents like this and others involving spamming, denial of service (DoS), virus attacks, content piracy, and malevolent hacking are becoming rampant. The security breaches that have posed a constant threat to desktop computers over the last decade are migrating to the world of wireless communications where they will pose a similar threat to mobile phones, smart phones, personal digital assistants (PDAs), laptop computers, and other yet-to-be-invented devices that capitalize on the convenience of wireless communications

SOLUTIONS

Standard

In 2003, the American College of Radiology (ACR) published a technical standard of teleradiology in which the DICOM standard (Bidgood & Horii, 1992) was used as a framework for medical-imaging communication. The DICOM standard was developed by the ACR and the National Electrical Manufacturers Association (NEMA) with input from various vendors, academia, and industry groups. Based upon the open system interconnect (OSI) reference model, which defines a seven-layer protocol, DICOM is an application-level standard, which means it exists inside layer 7. DICOM provides standardized formats for images, a common information model, application service definitions, and protocols for communication.

3G Network

3G stands for third generation (Collins & Smith, 2001) and is a wireless industry term for a collection of international standards and technologies aimed at increasing efficiency and improving the performance of mobile wireless networks (data speed, increased capacity for voice and data, and the advent of packet data networks vs. today's switched networks). As second-generation (2G) wireless networks evolve into third-generation systems around the globe, operators are working hard to enable 2G and 3G compatibility and worldwide roaming, including WCDMA, CDMA2000, UMTS, and EDGE technologies. In this project 3G technology was applied in teleradiology service for improving the speed of communication.

Types of 3G

Wideband Code Division Multiple Access (WCDMA)

This is a technology for wideband digital radio communications of Internet, multimedia, video, and other capacity-demanding applications. WCDMA has been selected for the third generation of mobile telephone systems in Europe, Japan, and the United States. Voice, images, data, and video are first converted to a narrowband digital radio signal. The signal is assigned a marker (spreading code) to distinguish it from the signal of other users. WCDMA uses variable rate techniques in digital processing and can achieve multi-rate transmissions. WCDMA has been adopted as a standard by the ITU under the name IMT-2000 direct spread.

Code Division Multiple Access 2000 (CDMA 2000)

Commercially introduced in 1995, CDMA quickly became one of the world's fastest-growing wireless technologies. In 1999, the International Telecommunications Union selected CDMA as the industry standard for new "third-generation" wireless systems. Many leading wireless carriers are now building or upgrading to 3G CDMA networks in order to provide more capacity for voice traffic, along with high-speed data capabilities. Today, over 100 million consumers worldwide rely on CDMA for clear, reliable voice communications and leading-edge data services.

Universal Mobile Telecommunication (UMTS)

This is the name for the third-generation mobile telephone standard in Europe, standardized by the European Telecommunications Standards Institute (ETSI). It uses WCDMA as the underlying standard. To differentiate UMTS from competing network technologies, UMTS is sometimes marketed as 3GSM, emphasizing the combination of the 3G nature of the technology and the GSM standard which it was designed to succeed. At the air interface level, UMTS itself is incompatible with GSM. UMTS phones sold in Europe (as of 2004) are UMTS/GSM dual-mode phones, hence they can also make and receive calls on regular GSM networks. If a UMTS customer travels to an area without UMTS coverage, a UMTS phone will automatically switch to GSM (roaming charges may apply). If the customer travels outside of UMTS coverage during a call, the call will be transparently handed off to available GSM coverage. However, regular GSM phones cannot be used on the UMTS networks.

Enhanced Data for Global Evolution (EDGE)

EDGE is a technology that gives GSM the capacity to handle services for the third generation of mobile telephony. EDGE was developed to enable the transmission of large amounts of data at a high speed, 384 kilobits per second. EDGE uses the same TDMA (time division multiple access) frame structure, logic channel, and 200 kHz carrier bandwidth as today's GSM networks which allows existing cell plans to remain intact.

Image Resolution

Digital images, whether viewed on a computer monitor, transmitted over a phone line, or stored on a hard disk or archival medium, are pictures that have a certain spatial resolution. The spatial resolution, or size, of a digital image is defined as a matrix with a certain number of pixels (information dots) across the width of the image and down the length of the image. The more pixels, the better the resolution. This matrix also has depth. This depth is usually measured in bits and is commonly known as shades of gray: a 6-bit image contains 64 shades of gray; 7-bit, 128 shades; 8-bit, 256 shades; and 12-bit, 4096 shades.

The size of a particular image is referenced by the number of horizontal pixels "by" (or "times") the number of vertical pixels, and then by indicating the number of bits in the shades of gray as the depth. For example, an image might have a resolution of 640 × 480 and 256 shades of gray, or 8 bits deep. The number of bits in the data set can be calculated by multiplying 640 × 480 × 8 equals 2,457,600 bits. Since there are 8 bits in a byte, the 640 × 480 image with 256 shades of gray is 307,200 bytes or .3072 megabytes of information.

Data Compression

Although images should be permanently archived as raw data or with only lossless data compression (no data is destroyed), hardware and software technologies exist that allow teleradiology systems to compress digital images into smaller file sizes so that the images can be transmitted faster. Compression is usually expressed as a ratio: 3:1, 10:1, or 15:1. The compression ratio refers to the ratio of the size of a compressed file to the original uncompressed file.

Certain images can withstand a substantial amount of compression without a visual difference: computed tomography and magnetic resonance images have large areas of black background surrounding the actual patient image information in virtually every slice. The loss of some of those pixels has no impact on the perceived quality of the image nor does it significantly change reader-interpretive performance.

Image Transmission

Image-transmission time is directly proportional to the file size of the digital image. The greater the amount of digital information in an image which involves the image matrix size and the number of bits per pixel, the longer the time required to transmit the image from one location to another. A radiological image contains a large amount of digital information. For example, an image with a relatively low resolution of 512 × 512 × 8 bits contains 2,097,152 bits of data, and a 1,024 × 1,024 × 8-bit image has 8,388,608 bits of data. Transmission time has to follow the laws of size. The only way to decrease the transmission time is either to increase the speed of the modem or reduce the number of bits (compress the image) being sent. The following formula is used to calculate the time to transmit an image:

Table 1. Shades of gray and matrix depth

Shades of Gray	Matrix Depth	Shades of Gray	Matrix Depth
256	8 bits	16	4 bits
128	7 bits	8	3 bits
64	6 bits	4	2 bits
32	5 bits	2	1 bit

(Matrix Size) × (Matrix Depth + 2 bits) × (Percentage of Compression) / (Modem Speed) = Seconds to Transmit

Matrix Depth is the shades of gray as shown in Table 1.

For modem or router control, most devices add 2 bits when transmitting as overhead.

Data Management

According to the ACR Technical Standard for Teleradiology, each examination data file must have an accurate corresponding patient and examination database record that includes patient name, identification number, examination date, type of examination, and facility at which the examination was performed. A Structural Query Language (SQL) database has been installed for the registration of each incoming and query of studies.

Web Technology

Web technology offers a significant advantage to physicians who need to receive images quickly, and who require real-time image navigation and manipulation to perform diagnostic tasks effectively. It facilitates the use of graphical-user interfaces, making teleradiology and picture achieving and communication system (PACS) applications easier to use and more responsive. Additionally, clients and servers can be run on different platforms, allowing end users to free themselves from particular proprietary architectures. Software applications designed for client-server computing can interface seamlessly with most hospital information systems (HISs) (RCR, 1999) or radiology information systems (RISs), while providing rapid soft-copy image distribution.

Storage

RAID (Marcus & Stern, 2003) stands for redundant array of inexpensive (or identical) disks. RAID employs a group of hard disks and a system that sorts and stores data in various forms to improve data-acquisition speed and provide improved data protection. To accomplish this, a system of levels (from 1 to 5) “mirrors,” “stripes,” and “duplexes” data onto a group of hard disks. All images were stored in the RAID of the server for high availability of the service.

Display

Today, most of the gray shades were produced by a mixing of primary colors in the video boards. Three types of video boards are commonly available, including 16-bit, 24-bit, and 32-bit color video cards. The video board with higher bits can be configured in a lower bit mode. The 16-bit color mode is called “High Color” mode with almost “good enough” quality to show photo images, at least for most purposes; 16-bit color is 5 bits each of red, green, and blue packed into one 16-bit word (2 bytes per pixel). Five bits can show 32 shades of each primary RGB channel, and 32 × 32 × 32 is 32K colors. Green used the extra one bit for 6 bits to achieve 64K colors overall, but half of them are green. The human eye is most sensitive to green-yellow, and more shades are a bigger advantage there. Green has twice the luminance of red and six times more than blue, so this is very reasonable. Video boards do vary, but 24 bits is normally not so much better in most cases, except in wide smooth gradients.

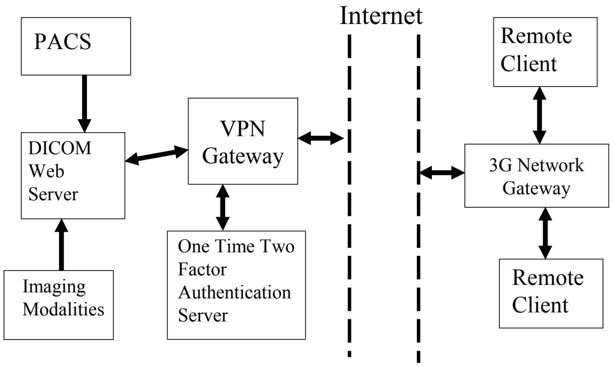
Video boards for the last few years are 24-bit color or “true color”; 24-bit color is 8 bits each of RGB, allowing 256 shades of each primary color, and 256 × 256 × 256 = 16.7 million color combinations. Studies show that the human eye can detect about 100 intensity steps (at any one current brightness adaptation of the iris), so 256 tones of each primary is more than enough. We would not see any difference between RGB (90,200,90) and (90,201,90), but we can detect 1% steps (90,202,90) (on a cathode ray tube (CRT) tube, but 18-bit LCD panels show 1.5% steps). So our video systems and printers simply do not need more than 24 bits.

Theoretically, there is no true 32-bit color display mode. The confusion is that 24-bit color mode normally uses 32-bit video mode today, referring to the efficient 32-bit accelerator chips (word size). The 24-bit color mode and so-called 32-bit video mode show the same 24-bit colors, the same 3 bytes RGB per pixel; 32-bit mode simply discards one of the four bytes (wasting 25% of video memory), because having 3 bytes per pixel severely limits video acceleration functions. Processor chips can only copy data in byte multiples (8, 16, 32, or 64 bits). A 24-bit copy done with a hardware video accelerator would require three 8-bit transfers per pixel instead of one 32-bit transfer; 32-bit video mode is for speed, and it shows 24-bit color.

Liquid Crystal Display (LCD)

In teleradiology, it is more convenient to use LCD for image display than CRT. In LCD, there are no CRTs. Instead, thin “sandwiches” of glass contain liquid-crystal filled cells (red, green, and blue cells) that make up a pixel. Arrays of thin film transistors (TFTs) provide the voltage power, causing the crystals to untwist and realign so that varying amounts of light can shine through each, creating images. This par-

Figure 1. Schematic diagram of 3G wireless medical image viewing system



ticular sensitivity to light makes LCD technology very useful in projection such as LCD front projectors, where light is focused through LCD chips.

Specifically, there are five layers to the LCD display: a backlight, polarized glass sheet, colored pixel layering, coating of liquid crystal solution that responds to signals off a wired grid of x and y coordinates, followed by a second glass sheet. To create an image, electrical charges, precision coordinated in various degrees and volts, effect the orientation of the liquid crystals, opening and closing them and changing the amount of light that passes through specific colors of pixels. LCD technology has increased its accuracy that can produce sharp and more accurate color images than earlier passive-matrix technologies.

One-Time and Two-Factor Authentication

OTTFA is an authentication protocol that requires two independent ways to establish identity and privileges in which at least one is continuously and non-repetitively changing. This contrasts with traditional password authentication, which requires only one factor such as the knowledge of a password in order to gain access to a system. OTTFA technique provides a secure authentication protocol for the teleradiology service.

Application Tunneling (Port Forwarding)

Application tunneling (or port forwarding) is a combination technique of routing by port combined with packet rewriting. A convention router examines the packet header and dispatches the packet on one of its other interfaces, depending on the packet destination address. Port forwarding examines the packet header and forwards it on to another host with the header rewriting depending on the destination port. The application of application tunneling is its inability of the destination machine to see the actual originator of the forwarded packets and instead seeing them as originating from the router. One of the applications of application tunneling is in a virtual private network gateway, as shown in Figure 1. Using the application tunneling technique, the security of the teleradiology system can be strengthened considerably.

High Available Server Cluster

The purpose of high available (HA) clustering is to maintain a non-stop teleradiology service for the users. In the current design, there are three image servers configured to form an HA cluster using a load-balancing switch. The user could access any one of the servers for making of diagnosis using the teleradiology technique. Other clustering techniques being used include operating system clustering, Internet protocol (IP) failover, and fault tolerance (FT) techniques.

Implementation Result

The overall design of the 3G wireless medical image viewing system is shown in Figure 1. The medical images received from the PACS or imaging modalities were stored in the DICOM Web server.

The operation of the system is shown in Figures 2 through 10. The users can use a laptop computer as a remote client for the connection to the Internet through a 3G network gateway provided by their service provider (see Figure 2). From the Internet, the users can make a connection to the VPN gateway and be authenticated by the OTTFA server. After login, the VPN gateway will redirect the user to the DICOM Web server using the application tunneling technique. After another login of the DICOM Web server, users can query the server for related studies. Finally, once the related studies are found and selected, all related images can be retrieved and displayed as shown in Figures 9 and 10.

Figure 2. A laptop computer for teleradiology

Figure 3. A One-Time Two-Factor Authentication device with a 3G phone for teleradiology

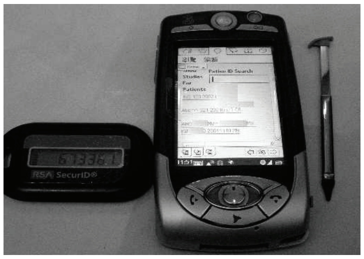
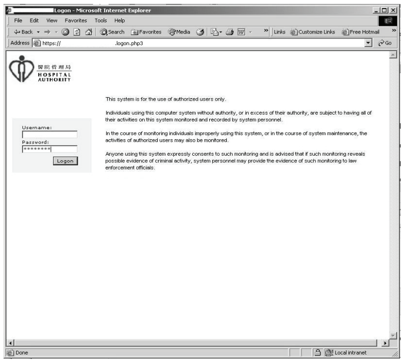

RSA SecurID

Figure 4. Login screen of VPN gateway


Logon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Links Customize Links Free Hotmail
Address https:// .logon.php3
Go
醫院管理局
HOSPITAL AUTHORITY
Logon
Done
Local intranet

This system is for the use of authorized users only.

Individuals using this computer system without authority, or in excess of their authority, are subject to having all of their activities on this system monitored and recorded by system personnel.

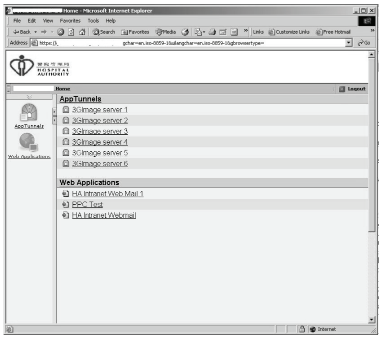
In the course of monitoring individuals improperly using this system, or in the course of system maintenance, the activities of authorized users may also be monitored.

Anyone using this system expressly consents to such monitoring and is advised that if such monitoring reveals possible evidence of criminal activity, system personnel may provide the evidence of such monitoring to law enforcement officials.

Username:

Password: ********

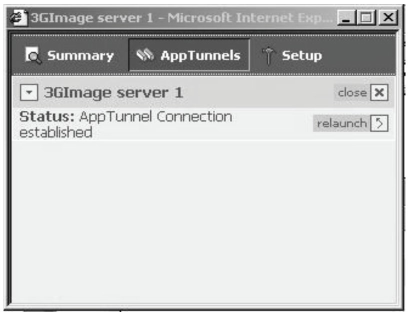
Figure 5. Successful login of VPN gateway


Home - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Links Customize Links Free Hotmail
Address https://
gchar=en.iso-8859-1&uilangchar=en.iso-8859-1&gbrowsertype=
Go
醫院管理局
HOSPITAL AUTHORITY
Home
Logout
AppTunnels
Web Applications
Internet

AppTunnels
- 3GImage server 1
- 3GImage server 2
- 3GImage server 3
- 3GImage server 4
- 3GImage server 5
- 3GImage server 6

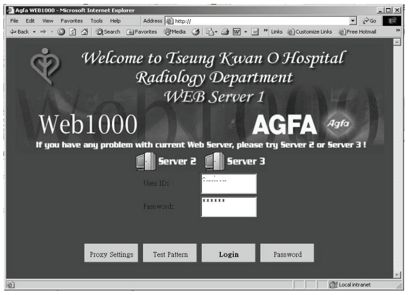
Web Applications
- HA Intranet Web Mail 1
- PPC Test
- HA Intranet Webmail

Figure 6. Application tunneling screen


3GImage server 1 - Microsoft Internet Exp...
Summary
AppTunnels
Setup
3GImage server 1
close
relaunch

Status: AppTunnel Connection established

Figure 7. Connection to DICOM Web server established


Agfa WEB1000 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://
Go
Back Search Favorites Media Links Customize Links Free Hotmail
Web1000
AGFA
Agfa
Server 2
Server 3
Proxy Settings
Test Pattern
Login
Password
Local intranet

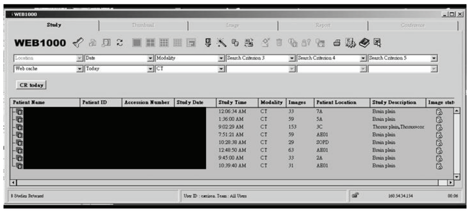
Welcome to Tseung Kwan O Hospital
Radiology Department
WEB Server 1

If you have any problem with current Web Server, please try Server 2 or Server 3 !

User ID:

Password: ******

Figure 8. Image data management


WEB1000
Study
Thumbnail
Image
Report
Conference
WEB1000
CR today

Location: Web cache
Date: Today
Modality: CT
Search Criterion 3:
Search Criterion 4:
Search Criterion 5:

Patient Name	Patient ID	Accession Number	Study Date	Study Time	Modality	Images	Patient Location	Study Description
[redacted]	[redacted]	[redacted]	[redacted]	12:06:34 AM	CT	33	7A	Brain plain
[redacted]	[redacted]	[redacted]	[redacted]	1:36:00 AM	CT	59	5A	Brain plain
[redacted]	[redacted]	[redacted]	[redacted]	9:02:29 AM	CT	153	3C	Thorax plain, Thorax+cor
[redacted]	[redacted]	[redacted]	[redacted]	7:51:21 AM	CT	59	AE01	Brain plain
[redacted]	[redacted]	[redacted]	[redacted]	10:28:38 AM	CT	29	SOPD	Brain plain
[redacted]	[redacted]	[redacted]	[redacted]	12:48:50 AM	CT	63	AE01	Brain plain
[redacted]	[redacted]	[redacted]	[redacted]	9:45:00 AM	CT	33	2A	Brain plain
[redacted]	[redacted]	[redacted]	[redacted]	10:39:40 AM	CT	31	AE01	Brain plain

8 Studies Returned | User ID : carrison. Team : All Users | 160.54.34.154 | 00:06

Figure 9. Selection of images

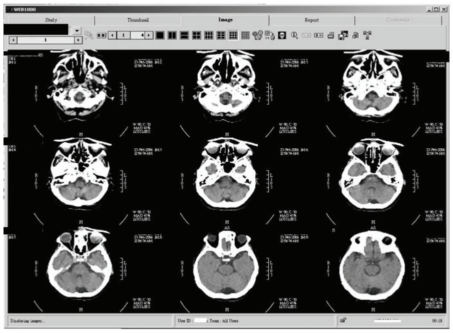
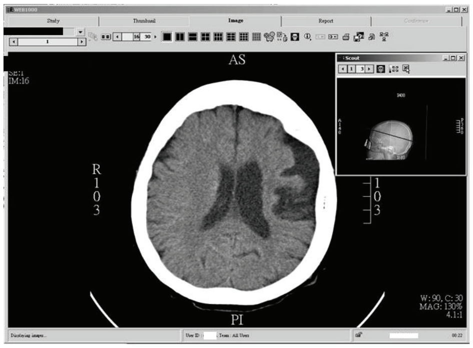

WEB1000
Study
Thumbnail
Image
Report
Conference
Displaying images...
Team : All Users

Figure 10. Viewing of images


WEB1000
Study
Thumbnail
Image
Report
Conference
SE:1
IM:16
AS
R
1
0
3
1
0
3
PI
W:90, C:30
MAG:130%
4.1:1
Scout
Displaying images...
Team : All Users

FUTURE TRENDS

4G Wireless Network

4G is the next generation of wireless networks that will replace 3G networks sometimes in future. In another context, 4G is simply an initiative by academic R&D laboratories to move beyond the limitations and problems of 3G to meet its promised performance and throughput. In reality, as of the first half of 2002, 4G is a conceptual framework for or a discussion point to address future needs of a universal high-speed wireless network that will interface with the wireline backbone network seamlessly. 4G also represents the hope and ideas of a group of researchers at Motorola, Qualcomm, Nokia, Ericsson, Sun, HP, NTT DoCoMo, and other infrastructure vendors who must respond to the needs of MMS, multimedia, and video applications if 3G never materializes in its full glory.

A comparison of key parameters of 4G with 3G is as shown in Table 2.

CONCLUSION

The above-mentioned 3G wireless medical image viewing system is providing a transfer speed of 384 kbps, which is comparable to a T1 fixed line of a speed of 1.4 mbps, but with greater accessibility and confidentiality. This system has been used successfully to ensure the availability of the teleradiology service 24 hours a day and 7 days a week. Throughout its construction and operation, it is found that the ACR and DICOM standards have provided useful guidelines on achieving the quality assurance and integrity expectations of this kind of service.

REFERENCES

ACR (American College of Radiology). (2003). *ACR technical standard for teleradiology.*

Bidgood, W. D., & Horii, S. C. (1992). Introduction to the ACRNEMA DICOM standard. *Radiographics, 12,* 34-35.

Collins, D., & Smith, C. (2001). *3G wireless networks.* New York: McGraw-Hill.

Marcus, E., & Stern, H. (2003). *Blueprints for high availability* (2nd ed.). New York: John Wiley & Sons.

Oguchi, K., Murase, S., Kaneko, T., Takizawa, M., & Kadoya, M. (2001). Preliminary experience of wireless teleradiology system using Personal Handyphone System. *Nippon Igaku Hoshasen Gakkai Zasshi, 61*(12), 686-687.

RCR (Royal College of Radiologists). (1999). *Guide to information technology in radiology: Teleradiology and PACS.* Board of Faculty of Clinical Radiology, RCR.

Reponen, J., Ilkko, E., Jyrkinen, L., Tervonen, O., Niinimaki, J., Karhula, V., et al. (2000). Initial experience with a wireless personal digital assistant as a teleradiology terminal for reporting emergency computerized tomography scans. *Journal of Telemed Telecare, 6*(1), 45-49.

Table 2. A comparison of 3G and 4G features

	3G (including 2.5G, sub3G)	**4G**
Major Requirement Driving Architecture	Predominantly voice driven—data was always add on	Converged data and voice over IP
Network Architecture	Wide area cell-based	Hybrid—integration of wireless LAN (WiFi, Bluetooth) and wide area
Speeds	384 Kbps to 2 Mbps	20 to 100 Mbps in mobile mode
Frequency Band	Dependent on country or continent (1800-2400 MHz)	Higher frequency bands (2-8 GHz)
Bandwidth	5-20 MHz	100 MHz (or more)
Switching Design Basis	Circuit and packet	All digital with packetized voice
Access Technologies	W-CDMA, 1xRTT, edge	OFDM and MC-CDMA (multi-carrier CDMA)
Forward Error Correction	Convolutional rate 1/2, 1/3	Concatenated coding scheme
Component Design	Optimized antenna design, multi-band adapters	Smarter antennas, software multiband and wideband radios
IP	A number of air link protocols, including IP 5.0	All IP (IP6.0)

Tong, C. K. S., Chan, K. K., & Wong, C. K. (2003). Common gateway interfacing and dynamic JPEG techniques for remote handheld medical image viewing. *CARS,* 815-820.

KEY TERMS

Application Tunneling: A combination technique of routing by port combined with packet rewriting.

Digital Imaging and Communication in Medicine (DICOM): A standard developed by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) to provide standardized formats for images, a common information model, application service definitions, and protocols for communication.

One-Time Two-Factor Authentication: An authentication protocol that requires two independent ways to establish identity and privileges in which at least one is continuously and non-repetitively changing.

Port Forwarding: Another name for application tunneling.

Redundant Array of Inexpensive (or Identical) Disks (RAID): Employs a group of hard disks and a system that sorts and stores data in various forms to improve data-acquisition speed and provide improved data protection.

Teleradiology: The process of sending radiologic images from one point to another through digital, computer-assisted transmission, typically over standard telephone lines, a wide-area network (WAN), or over a local area network (LAN).

Mobile Multicast

Thomas C. Schmidt
HAW Hamburg, Germany

Matthias Wählisch
FHTW Berlin, Germany

INTRODUCTION

The submission of datagrams from a node to a group of receivers must be seen as a powerful extension of the Internet routing layer (Deering, 1989). Multicast group communication forms an integral building block of a wide variety of applications, ranging from public content distribution and streaming over voice and videoconferencing, collaborative environments, and gaming up to the self-organization of distributed systems. Its support by network layer multicast will be needed, whenever globally distributed, scalable, serverless, or instantaneous communication is required. As broadband media delivery more and more emerges to be a typical mass scenario, scalability and bandwidth efficiency of multicast routing continuously gains relevance. Internet multicasting will be of particular importance to mobile environments, where users commonly share frequency bands of limited capacity. The rapidly increasing mobile reception of 'infotainment' streams may soon require a wide deployment of mobile multicast services.

The fundamental approach to deal with mobility in the next-generation Internet is stated in the Mobile IPv6 RFC (Johnson, Perkins, & Arkko, 2004). Multicast has only roughly been treated therein, but raises quite distinctive aspects within a mobility-aware Internet infrastructure. On the one hand multicast routing itself supports dynamic route configuration, as members may join and leave ongoing group communication over time. On the other hand multicast group membership management and routing procedures are intricate and too slow to function smoothly for mobile users. In addition multicast imposes a special focus on source addresses. Applications commonly identify contributing streams through source addresses, which must not change during sessions, and routing paths in most protocols are chosen from destination to source.

Mobile multicast has been a concern for about ten years (Xylomenos & Polyzos, 1997) and led to innumerous proposals for solutions. Intricate multicast routing procedures, though, are not easily extensible to comply with mobility requirements. Any client subscribed to a group while in motion requires delivery branches to pursue its new location; any mobile source requests the entire delivery tree to adapt to its changing positions. Significant effort has been already invested in protocol designs for mobile multicast receivers. Only limited work has been dedicated to multicast source mobility, which poses the more delicate problem (Romdhani, Kellil, Lach, Bouabdallah, & Bettahar, 2004a). In multimedia conference scenarios each member commonly operates as receiver and as sender for multicast-based group communication. In addition, real-time communication such as voice or video over IP places a severe temporal requirement on mobility protocols: seamless handover scenarios need to limit disruptions or delay to less than 100 milliseconds. Jitter disturbances are not to exceed 50 milliseconds. Note that 100 milliseconds is about the duration of a spoken syllable in real-time audio.

While multicast routing itself has been proposed to support mobility on the Internet (Helmy, 2000), consensus on an efficient, robust, and widely deployable scheme of multicast for mobile hosts is still lacking (Schmidt & Wählisch, 2005a). In this review we will summarize the state of the art in current work to multicast to Mobile IPv6 networks. The principle conceptual problems are discussed and analyzed. Propositions for improvement and possible directions to further proceed towards a mobile multicast are presented.

BACKGROUND

Multicast mobility must be considered a generic term that subsumes a collection of quite distinct functions. First, multicast communication divides into any source multicast (ASM) (Deering, 1989) and source-specific multicast (SSM) (Bhattacharyya, 2003; Holbrook & Cain, 2005). Second, the roles of senders and receivers are asymmetric and need distinction. Both individually may be mobile. Their interaction is facilitated by a multicast routing function—such as DVMRP (Waitzman, Partridge, & Deering, 1988), PIM-SM/SSM (Estrin et al., 1998) or CBT (Ballardie, 1997)—and the IPv6 multicast listener discovery protocol (Deering, Fenner, & Haberman, 1999).

Any multicast mobility solution must account for all of these functional blocks. It should enable seamless continuity of multicast sessions when moving from one IPv6 subnet to another. It should preserve the multicast nature of packet distribution and approximate optimal routing. It should sup-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

port per-flow handover for multicast traffic, as properties and designations of flows may be of individual kind.

Multicast routing dynamically adapts to session topologies, which then may change under mobility. However, routing convergence arrives at a time scale of seconds, even minutes and is far too slow to support seamless handovers for interactive or real-time media sessions. The actual temporal behavior strongly depends on the routing protocol in use and on the geometry of the current distribution tree. A mobility scheme that arranges for adjustments—that is, partial changes or full reconstruction—of multicast trees is forced to make provision for time buffers sufficient for protocol convergence. Special attention is needed with a possible rapid movement of the mobile node, as this may occur at much higher rates than compatible with protocol convergence.

IP layer multicast packet distribution is an unreliable service, which is bound to connectionless transport protocols. Packet loss thus will not be handled in a predetermined fashion. Mobile multicast handovers should not cause significant packet drops. Due to statelessness the bi-casting of multicast flows does not cause foreseeable degradations of the transport layer.

Group addresses in general are location transparent, even though there are proposals to embed unicast prefixes or rendezvous point addresses (Savola & Haberman, 2004). Source addresses contributing to a multicast session are interpreted by the routing infrastructure and by receiver applications, which frequently are source address aware. Multicast therefore inherits the mobility address duality problem for source addresses, being a logical node identifier (HoA) on the one hand and a topological locator (CoA) on the other.

Multicast sources in general operate decoupled from their receivers in the following sense: a multicast source submits data to a group of unknown receivers, thus operating without any feedback channel. It neither has means to inquire on properties of its delivery trees, nor will it be able to learn about the state of its receivers. In the event of an inter-tree handover, a mobile multicast source therefore is vulnerable to losing receivers without taking notice.

MULTICAST LISTENER MOBILITY

A mobile multicast listener entering a new IP subnet may simply re-subscribe to its previously received groups. It thereby will encounter one of the following conditions. In the new network the requested multicast service may be supported, but the multicast groups under subscription may not be forwarded to it. The current distribution trees for the desired groups may reside at large routing distance. It may as well occur that some or all groups under subscription of the mobile node are received by one or several local group members at the instance of arrival and that multicast streams natively flow.

The temporal behavior of the multicast handover will largely depend on the conditions met. The mobile node may employ predictive or reactive accelerating schemes to reduce performance degradation. For a detailed analysis, see Schmidt and Wählisch (2005b).

ASM SOURCE MOBILITY

A node submitting data to an Any Source Multicast group defines the root of either a shared or a source-specific delivery tree. Beside root location forwarding along, this delivery tree will be bound to a topological network address due to reverse path forwarding (RPF) checks. A mobile multicast source moving away is *solely* enabled to either inject data into a previously established delivery tree by using its previous topologically correct source address, or to (re-)define a multicast distribution tree compliant to its new location. In pursuing the latter the mobile sender will have to proceed without control of the new tree construction due to decoupling of sender and receivers.

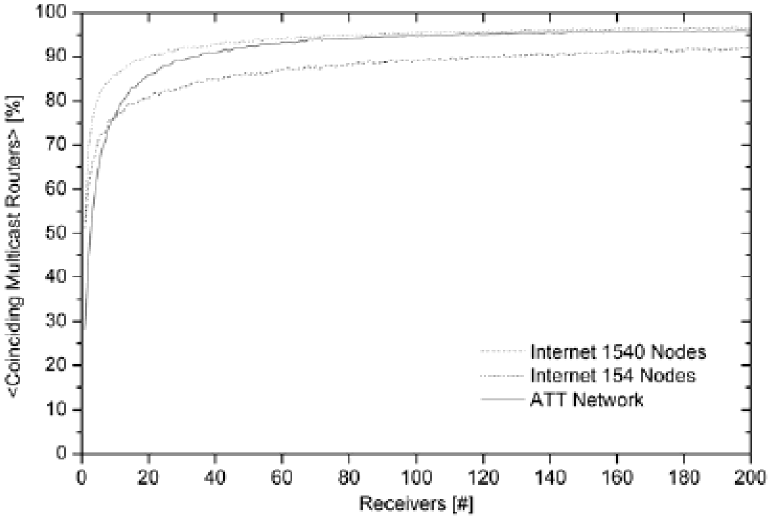
Conforming to address transparency and temporal handover constraints will be the key problems for any route optimizing mobility solution. Additional issues arrive from possible packet loss and from multicast scoping. A mobile source away from home must attend scoping restrictions which arise from its home and its visited location (Johnson et al., 2004).

SSM SOURCE MOBILITY

Fundamentally Source Specific Multicast has been designed for changeless addresses of multicast senders. Source addresses in client subscription to SSM groups are directly used for route identification. Any SSM subscriber is thus forced to know the topological address of its group contributors. SSM source identification invalidates when source addresses change under mobility.

Consequently, source mobility for SSM packet distribution introduces a significant conceptual complexity in addition to the problems of mobile ASM. As a listener is subscribed to an (S,G) channel membership and as routers have established an (S,G)-state shortest path tree rooted at source S, any change of source addresses under mobility requests for state updates at all routers and all receivers. A moving source would have to update its change of CoA with all listeners, which subsequently had to newly subscribe and initiate corresponding source-specific trees. As the principle multicast decoupling of a sender from its receivers likewise holds for SSM, the need for client update turns into a severe problem.

Figure 1. Relative router coincidence between subsequent multicast distribution trees rooted at 5 hops distance



PROPERTIES OF MULTICAST DISTRIBUTION TREES

The efficiency of multicast distribution trees has been studied well and was empirically derived to follow a Chuang and Sirbu (1998) scaling law—that is, the number of links employed in a multicast tree with m receivers is well approximated by $<L_u> m^{0.8}$, where $<L_u>$ represents the average number of unicast hops for the corresponding network. Van Mieghem, Hooghiemstra, and van der Hofstad (2001) semi-analytically derived that the number of routers in the Internet, which can be reached from a root, grows exponentially in the number of hops with an effective router degree of 3.2. From the perspective of mobile sources, the geometric evolution of multicast trees according to a moving root is of increased importance. The number of hops between subsequent root routers, the mobility "step-size," should serve as an appropriate measure of variation.

Source-specific shortest path trees subsequently generated from mobility steps are highly correlated. They most likely branch towards identical receivers and are rooted at a distance, which corresponds to the "step size" undertaken by the moving source. Figure 1 visualizes the relative change of distribution trees as a function of receiver multiplicity for a medium step size of 5. It is interesting to note that even in larger networks, a range from 80 to above 90% of multicast tree routers remains fixed under a mobility step. Simulations have been performed based on real-world Internet topology data, representing the ATT core network (Heckmann, Piringer, Schmitt, & Steinmetz, 2003) and subsamples from the SCAN Project (2005).

SOLUTIONS

Two principal approaches to mobile multicast have been foreseen in (Johnson et al., 2004):

Remote subscription forces the mobile node to re-initiate multicast distribution subsequent to handover, using its current Care-of Address. This approach of tree discontinuation relies on the dynamics of multicast routing to adapt to network changes. Aside from service disruption on handover, remote subscription allows for a transparent mobility management for mobile listeners. Mobile ASM senders will experience a mobility-driven change of source addresses, and thus need to transmit a home address destination option to maintain session persistence at the application layer. Mobile SSM sources cannot follow a pure remote subscription approach, as receivers are subscribed to a static source address and thus will lose communication on handovers.

Bi-directional tunneling guides the mobile node to tunnel all multicast data via its home agent. This solution hides all movement and results in static multicast trees. It transparently may be employed by mobile multicast listeners and sources, on the price of triangular routing and possibly significant performance degradations due to widely spanned data tunnels. The additional delay due to packet forwarding through the home agent strongly depends on network topology and has been evaluated in Schmidt and Wählisch (2006). For densely meshed single provider networks, an average delay excess of about 50% was found, whereas in large samples of 15,000 inter-provider core nodes, the bi-directional tunneling more than doubled network transmission times.

Both schemes carry the distinct advantage of immediate employability, as they do not require changes to existing protocols. Accordingly the combination of operating remote subscription at mobile receivers and bi-directional tunneling at mobile sources (i.e., hybrid approaches) can be frequently found in the literature (Romdhani et al., 2004a).

Agent-based inter-tree handovers have been established as the common approach to compensate for performance deficits in remote subscription or bi-directional tunneling. They attempt to balance between the anchorless remote subscription and the purely static bi-directional tunneling. Static agents typically act as local tunneling proxies, allowing for some inter-agent handover while the mobile node moves away. A decelerated inter-tree handover will be the outcome of agent-based multicast mobility, where extra effort is needed to sustain session persistence through address transparency of mobile sources.

Aside from many conceptual papers (Romdhani et al., 2004a), there are proposals of agent-based approaches compliant to the unicast real-time mobility infrastructure of Fast MIPv6, the M-FMIPv6 (Suh, Kwon, Suh, & Park, 2004), and of Hierarchical MIPv6, the M-HMIPv6 (Schmidt & Wählisch, 2005c), and to context transfer (Jonas & Miloucheva, 2005).

"Fast Multicast Protocol for Mobile IPv6 in the Fast Handover Environments" adds support for mobile multicast receivers to Fast MIPv6. On predicting a handover to a next access router, the mobile node submits its multicast group addresses under subscription with its fast binding update to the previous access router. Routers thereafter exchange those groups. In the ideal case, the new access router will be enabled to subscribe to all requested groups, even before the MN has disconnected from its previous network.

"Seamless Multicast Handovers in a Hierarchical Mobile IPv6 Environment (M-HMIPv6)" extends the hierarchical MIPv6 architecture to support mobile multicast receivers and sources. Mobility anchor points act as proxy home agents, controlling group membership for multicast listeners and issuing traffic to the network in place of mobile senders. Handovers within a domain of one mobility anchor point remain invisible in this micro mobility approach. At the event of an inter-anchor handover, the previous anchor point will be triggered by a reactive binding update and act as a proxy forwarder. Subsequent to MIPv6 handover, continuous data reception is thus assured, while a remote subscription continues within the new domain. A home address destination option has been added to streams from a mobile sender. Consequently transparent source addressing is provided to the socket layer.

It should be noted that none of the above approaches addresses SSM source mobility, except the bi-directional tunneling. Jelger and Noel (2002) suggest mobile SSM handover improvements by employing anchor points within the source network, causing a continuous data reception during client-initiated handovers. Their approach can be understood as a fairly direct transformation of traditional agent-based solutions in ASM mobility: in the event of a handover, the mobile source reconnects with the previous anchor point and distributes a new "SSM-Source Handover Notification" binding update option. On reception of this binding update, listeners re-subscribe to the new source address channel, while multicast streams continuously are bi-casted down the previous distribution tree. In applying an ASM-type agent-based approach, the authors disregard consequences of the decoupling principle for SSM. Bi-casting to the previous SSM tree cannot be terminated until all receivers have established membership of the new distribution tree, which is of unforeseeable progress.

Tree modification schemes have deserved very little attention as procedures that modify existing distribution trees to continuously serve for data transmission of mobile sources. In the case of DVMRP routing, Chang and Yen (2004) propose an algorithm to extend the root of a given delivery tree to incorporate a new source location in ASM. To fix DVMRP forwarding states and heal RPF-check failures, the authors rely on a complex additional signaling protocol.

Focusing on inter-domain mobile multicast routing in PIM-SM, Romdhani et al. (2004b) propose a tunnel-based backbone distribution of packets between newly introduced "mobility-aware rendezvous points" (MRPs). These MRPs operate on extended multicast routing tables, which simultaneously hold HoA and CoA. This solution accounts for the ASM inter-domain source activation problem (Romdhani et al., 2004a).

O'Neill (2002) suggests a scheme to overcome reverse path forwarding (RPF) check failures originating from multicast source address changes, by introducing extended routing information, which accompanies data in a hop-by-hop option header.

Finally Schmidt and Wählisch (2006) introduce a scheme for mobile SSM sources of continuously morphing the previous distribution tree into the new tree. A mobile multicast source (MS) away from home will transmit unencapsulated data to a group using its HoA on the application layer and its current CoA on the Internet layer, just as unicast packets are transmitted by MIPv6. In extension to unicast routing, though, the entire Internet layer (i.e., routers included) will be aware of the permanent HoA. Subsequent to handover the mobile source will immediately continue to deliver data along an extension of its previous source tree. Delivery is done by elongating the root of the previous tree from the previous designated router to the next designated router. All routers along the path, located at root elongation or previous delivery tree, thereby will learn MS's new CoA and implement appropriate forwarding states. Routers on this extended tree will use RPF checks to discover potential shortcuts. Registering a new CoA as source address, those routers that receive the state update via the topologically incorrect interface will submit a join in the direction of a new shortest path tree and prune the old tree membership as soon as data arrives. All other routers will re-use those parts of the previous delivery tree that which coincide with the new shortest path tree.

Tree modification schemes introduce the potential of unencapsulated data transmission to multicast groups continuous under mobility. They are well suited for ASM and SSM source mobility, and minimize packet loss and delay. However, their disadvantage lies in the requirement to change existing multicast routing protocols.

FUTURE TRENDS

During the past decade multicast routing has encountered hesitance in deployment, which is mainly due to its complexity and potential threads as the outcome of unrestricted packet duplication in ASM. It is widely believed that simpler mechanisms for group distribution in Source Specific Multicast will lead to a pervasive dissemination of multicast infrastructure and services throughout the Internet. Facing mobility, a significant extra burden is added onto multicast deployment, which is even enhanced for SSM. The dispute on multicast network efficiency vs. deployment complexity thus arouses anew (Garyfalos, Almeroth, & Sanzgiri, 2004).

However, multicast services in a mobile environment may soon become inescapable, when multimedia distribution services will develop as a strong business case for IP portables. At first, the deployment of simple bi-directional tunneling or micromobile proxy services can be foreseen. As mobility will unfold dominance and as efficiency will show a larger impact in costly radio environments, the evolution of multicast protocols will naturally follow mobility constraints. This future-generation Internet routing will have to provide seamless mobile multicasting and presumably will operate directly on the shapes of distribution trees.

CONCLUSION

Mobile multimedia group services are one of the major driving forces, but also a severe challenge for the Internet infrastructure today. In this overview we discussed the principle conceptual difficulties of mobile multicast and derived the current state of the art for solutions. A special focus has been donated to source mobility from the perspective of Source Specific Multicast. Initial ideas for new mobile SSM communication and routing schemes were presented, as they may serve as a starting point for revitalizing the research on practically feasible solutions to this intricate field of development.

REFERENCES

Ballardie, A. (1997). *Core Based Trees (CBT version 2) multicast routing.* RFC 2189, IETF.

Bhattacharyya, S. (2003). *An overview of Source-Specific Multicast (SSM).* RFC 3569, IETF.

Chang, R.-S., &. Yen, Y.-S (2004). A multicast routing protocol with dynamic tree adjustment for Mobile IPv6. *Journal of Information Science and Engineering, 20,* 1109-1124.

Chuang, J., & Sirbu, M. (1998, July). Pricing multicast communication: A cost-based approach. *Presented at INET'98,* Geneva, Switzerland.

Deering, S. (1989). *Host extensions for IP multicasting.* RFC 1112, IETF.

Deering, S., Fenner, W., & Haberman, B. (1999). *Multicast Listener Discovery (MLD) for IPv6.* RFC 2710, IETF.

Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P., & Wei, L. (1998). *Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol specification.* RFC 2362, IETF.

Garyfalos, A., Almeroth, K., & Sanzgiri, K. (2004). Deployment complexity versus performance efficiency in mobile multicast. In *Proceedings of Broadnets '04,* (published online).

Heckmann, O., Piringer, M., Schmitt, J., & Steinmetz, R. (2003). On realistic network topologies for simulation. In *MoMe-Tools '03: Proceedings of the ACM SIGCOMM Workshop on Models, Methods and Tools for Reproducible Network Research* (pp. 28-32).

Helmy, A. (2000). A multicast-based protocol for IP mobility support. In *Proceedings of the 2^nd^ International Workshop of Networked Group Communication (NGC2000)* (pp. 49-58). New York: ACM Press.

Holbrook, H., & Cain, B. (2005, October 7). *Source-specific multicast for IP.* IETF Internet draft, work in progress.

Jelger, C., & Noel, T. (2002, January). *Supporting mobile SSM sources for IPv6 (MSSMSv6).* IETF Internet draft, work in progress.

Johnson, D. B., Perkins, C., & Arkko, J. (2004). *Mobility support in IPv6.* RFC 3775, IETF.

Jonas, K., & Miloucheva, I. (2005, June). *Multicast context transfer in Mobile IPv6.* IETF Internet draft, work in progress.

O'Neill, A. (2002, July). *Mobility management and IP multicast.* IETF Internet draft, work in progress.

Romdhani, I., Kellil, M., Lach, H.-Y., Bouabdallah, A., & Bettahar, H. (2004a). IP mobile multicast: Challenges and solutions. *IEEE Communication Surveys & Tutorials, 6*(1), 18-41.

Romdhani, I., Kellil, M., Lach, H.-Y., Bouabdallah, A., & Bettahar, H. (2004b). Mobility-aware rendezvous point for mobile multicast sources. In *Proceedings of WWIC 2004* (LNCS 2957, pp. 62-73). Berlin: Springer-Verlag.

Savola, P., & Haberman, B. (2004). *Embedding the Rendezvous Point (RP) address in an IPv6 multicast address.* RFC 3956, IETF.

SCAN Project. (2005). *Internet maps. SCAN+Lucent map.* Retrieved from http://www.isi.edu/scan/mercator/maps.html

Schmidt, T. C., & Wählisch, M. (2005a, October). *Multicast mobility in MIPv6: Problem statement.* IRTF Internet draft, work in progress.

Schmidt, T. C., & Wählisch, M. (2005b). Predictive versus reactive—Analysis of handover performance and its implications on IPv6 and multicast mobility. *Telecommunication Systems, 30*(1-3), 123-142.

Schmidt, T. C., & Wählisch, M. (2005c, December). *Seamless multicast handover in a hierarchical Mobile IPv6 environment (M-HMIPv6).* IETF Internet draft, work in progress.

Schmidt, T. C., & Wählisch, M. (2006, April). A first performance analysis of the tree morphing approach to source mobility in source specific multicast routing. In *Proceedings of IEEE ICN'06.* IEEE Press.

Suh, K., Kwon, D.-H., Suh, Y.-J., & Park, Y. (2004, February). *Fast multicast protocol for Mobile IPv6 in the fast handovers environments.* Internet draft, work in progress.

Xylomenos, G., & Polyzos, G. C. (1997). IP multicast for mobile hosts. *IEEE Communications Magazine*, *35*(1), 54-58.

Van Mieghem, P., Hooghiemstra, G., & van der Hofstad, R. (2001). On the efficiency of multicast. *IEEE/ACM Transactions on Networking, 9*(6), 719-732.

Waitzman, D., Partridge, C., & Deering, S. (1988). *Distance vector multicast routing protocol.* RFC 1075, IETF.

KEY TERMS

Any Source Multicast (ASM): The distribution of packets from an arbitrary source to a group of receivers. Receivers are addressed by a delocalized group address and remain unidentified by the source.

Bi-Casting: A technique of duplicating packets on the network to reduce packet loss. Forwarding, for example, may be done via new and previous access routers.

Care-of Address (CoA): The mobile node's temporary IP address in its visited network.

Home Address (HoA): The mobile node's permanent IP address in its home network.

Home Agent (HA): A router representing the absent mobile node in its home network. It intercepts packets destined to the mobile node's home address and tunnels these to the Care-of Address of the mobile node.

Inter-Tree Handover: Change of packet distribution from one to another multicast delivery tree.

Rendezvous Point: The root of a per-group *shared tree.*

Reverse Path Forwarding (RPF) Check: A routing algorithm verifying whether a multicast packet was received on the closest interface to the root of the distribution tree. RPF checks prevent multicast packet distribution from exhibiting loops.

Shared Tree: A multicast distribution tree rooted at a source-independent *rendezvous point* and serves for packet distribution of different of multiple sources.

Source-Specific Multicast (SSM): The distribution of packets from an initially specified source to a group of receivers. Listeners must actively subscribe by address to each source they want to receive traffic from.

Source-Specific Tree: A multicast distribution tree rooted at the multicast sender.

Mobile Payment and the Charging of Mobile Services

Key Pousttchi
University of Augsburg, Germany

Dietmar Georg Wiedemann
University of Augsburg, Germany

INTRODUCTION

According to the sweeping enthusiasm that characterized much of the news reporting in the years 1999 and 2000, mobile phones should by now have been firmly established as payment terminals in the most diverse fields. However, reality today is a different matter. Mobile payment as an established payment system seems to be a distant prospect in the case of most countries.

Since the mid-1990s there have been serious efforts to use mobile phones for payment processes. The starting point for these considerations was the fact that mobile phones are particularly suitable for conducting payment processes due to their specific characteristics, high diffusion in population, and users' positive attitude towards them (e.g., Henkel, 2002). In recent years several studies showed that customers in principle take an interest in mobile payment (e.g., Khodawandi, Pousttchi, & Wiedemann, 2003; Eisenmann, Linck, & Pousttchi, 2004). A further study bridged the gap of these and other studies' explanatory power, and confirmed a high interest also in the total population. During a representative study in September 2004, 49.6% of the German participants indicated that they are interested in and willing to use mobile payment (MobilMedia, 2004).

The commercial history of mobile payment procedures is short, but simultaneously characterized by rapid development. One of the first commercial mobile payment procedures was launched by the Finnish *mobile network operator* (MNO) Sonera in 1997. Customers were able to pay for goods at vending machines (Dahlberg, Mallat, & Öörni, 2003). New technological innovations used in mobile payment procedures and new use case scenarios for mobile payment have been developed at an increasingly fast pace ever since. Among the leading countries in mobile payment are Austria, South Korea, Singapore, Norway, Spain, Japan, Finland, and Italy, in which end-to-end solutions and clear business models have proved to be sustainable after four to five years of field trials and pilot projects (Taga & Karlsson, 2004).

However, in other countries the situation is disillusioning. For instance on the German market (which is not only the most important European market, but also a good sample for developments in many western markets), banks (e.g., Payitmobile), MNOs (e.g., Genion M-Payment), as well as quite a number of specialized intermediaries (e.g., Paybox, Geldhandy, or Street Cash) tried one's luck in recent years. Also the vertical alliance of the four large-scale and internationally active MNOs—Orange, Telefonica Moviles, T-Mobile, and Vodafone—was not able to start its integrated mobile payment system Simpay. When it was initiated in 2002, the primary objective was to introduce a pan-European mobile payment system for all payment scenarios. However, after six months a smaller compromise was made: providing a solution for their most urgent problem, charging mobile services, and additionally enabling payments for digital goods on the Internet. Also this did not come off. After numerous delays and intestine strife between the founders, Simpay finally stopped its activities in the middle of 2005 (Pousttchi & Wiedemann, 2006).

Thus, it can be concluded for the majority of countries that most mobile payment procedures were quit after the test stage, and procedures that came into the market had some diffusion, but outside of Asia, not many of these can be categorized as economically successful, although the preconditions for acceptance of mobile payment by customers are very good.

BACKGROUND

The diffusion of mobile phones during the nineties and the success of mobile services such as ring tones and logos have raised high expectations toward mobile commerce. We define *mobile commerce* as any kind of business transaction, on the condition that at least one side uses mobile communication techniques (Turowski & Pousttchi, 2004).

Mobile payments are expected to become one of the most important applications in mobile commerce (Varshney & Vetter, 2002). On closer examination of mobile payment, we have to differentiate two basic functions: payments inside and outside mobile commerce. Inside mobile commerce mobile payment is used for payments of mobile offers and is ideally system inherent. In the area of charging mobile services, we distinguish two basic terms: mobile billing and mobile payment. We refer to *mobile billing* as billing of

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

telecommunication services by an MNO within an existing billing relationship (Turowski & Pousttchi, 2004). The MNO could also be a mobile virtual network operator (MVNO) to which our following models and concepts would apply analogously. We define *mobile payment* as that type of payment transaction processing in the course of which—within an electronic procedure—(at least) the payer employs mobile communication techniques in conjunction with mobile devices for initiation, authorization, or realization of payment (Pousttchi, 2003). If a mobile payment procedure is provided by an MNO, we will have the intersection of mobile billing and mobile payment.

Analyzing different mobile payment procedures on joint characteristics, Kreyer, Pousttchi, and Turowski (2002) derive five standard types. The standard type with the most practical relevance today is the standard type phone bill, which is characterized by an MNO as the mobile payment service provider and the mobile phone bill as the settlement method. These procedures are normally either limited to the mobile commerce scenario or especially developed for this scenario. An example for the first case is the system inherent payment procedure of i-mode; examples for the second case are the different applications of premium rate SMS and the procedure m-pay of Vodafone.

As later discussed, mobile payment is crucial for mobile commerce, but not limited to this scenario. Outside mobile commerce, a mobile payment procedure can be understood as a mobile commerce application to complete payments in different situations. For this purpose four general settings, defined as payment scenarios, are to be considered (Kreyer et al., 2002; Khodawandi et al., 2003): transaction on the stationary Internet (electronic commerce scenario), at any kind of vending machine (stationary merchant automat scenario), in traditional retail (stationary merchant person scenario), and between end-customers (customer-to-customer scenario). The emphasis of this article is on mobile payment inside mobile commerce (mobile commerce scenario).

Analyzing the business model of a mobile service, we can distinguish—similar to electronic commerce—between direct and indirect revenue sources and transaction-dependent and transaction-independent revenue types (Turowski & Pousttchi, 2004, according to Wirtz, 2001). Concepts based exclusively on indirect revenues, for example, financed by advertisement, already failed on the stationary Internet, except for very few exceptions. The realization of transaction-independent revenues in mobile commerce (e.g., by sale of a subscription) is appropriate for certain kinds of services. However, subscription will have a rather inhibiting effect on the diffusion of many typical mobile services, in particular if customers want to use the service spontaneously or only occasionally. If direct transaction-dependent revenues are to be realized, then an adequate charging form between providers and customers will be necessary. Whereas in electronic commerce we still see an important role of traditional payment systems (e.g., Krueger, Leibold, & Smasal, 2006), a payment system for mobile commerce will typically not be adequate until it shares fundamental characteristics of the mobile offer it is to bill for, in particular its ubiquity (Pousttchi, Selk, & Turowski, 2002). This is in line with Coursaris and Hassanein (2002) and Mallat (2004). Their arguments are based on the fact that mobile commerce provides an opportunity for customers to reach services anytime and anywhere, and this implicates that also the payment procedure needs to follow these properties. Likewise, already Kieser (2001) Zobel (2001) derived the necessity of available mobile payment procedures from the necessity of charging of goods and services in the mobile commerce scenario, and additionally from the fact that traditional payment procedures are inapplicable in mobile commerce. Since companies are not going to invest in the development of innovative mobile applications or services unless they can be charged for appropriately, the existence of adequate possibilities for charging of the goods and services is crucial (Pousttchi et al., 2002). This is in line with Dahlberg et al. (2003) who stated that problems with payments profoundly hinder the development of mobile commerce.

As a result mobile payment is crucial for, but not limited to the mobile scenario. On the contrary, the universal applicableness of a mobile procedure in scenarios outside mobile commerce is relevant for its acceptance (Kreyer et al., 2002).

MOBILE PAYMENT INSIDE MOBILE COMMERCE

Offer Models

The most important subset of mobile commerce is the area of mobile value-added services. Due to the fact that the transmission of data is a substantial component, it is classified as a telecommunication service in the broader sense. Hence, the charging by the MNO is legally allowed in most countries, whereas payment outside of mobile commerce often requires a banking license. Typical examples for mobile added-value services are news, financial information services, or entertainment services. In principle, we distinguish two offer models: the offer by the MNO and the direct offer by a mobile content provider (Turowski & Pousttchi, 2004). The offer by the MNO is an MNO-centered solution. The MNO produces mobile services or buys them from a mobile content provider (just as he or she buys into network infrastructure or mobile devices) and thus offers a single face to the customer for network and services. An explicit payment process is not necessary, because typically only transmission is charged and consequently, mobile billing is applied similar to mobile voice services. This model was (and still is) usual on many markets and documents the market power of the

MNO inhibiting the direct relationship between customers and mobile content providers.

However, this poses a couple of problems to the MNO. Procurement and offer of content are not counted among MNO core competencies, whereupon these are very complex issues, especially if successful offers are to be made for different target groups. But the effort's result is mostly only an increased data transmission and consequently improved network capacity utilization. Thus, the offer of high-quality content pays off only conditionally through volume-dependent pricing. To come up against this, it is also possible to demand additional remuneration for high-quality content. However, the MNO relatively hopelessly competes with other actors counting content among their core competence—that is, branded content providers with an existing customer base like media companies or major sports clubs—and often already possesses an existing customer relationship outside of mobile commerce.

This is the reason why in times of 2.5G and 3G mobile networks, the model direct offer by the mobile content provider following the i-mode paradigm from Japan will prevail to an increasing degree (Natsuno, 2003). In this model, the mobile content provider has a direct customer relationship and generates direct revenues by the service offer. By means of service's content and quality, the mobile content provider provides customers with added value that is paid by them in addition to the transmission. The mobile content provider and MNO compensate added value and transmission costs. In this model the mobile content provider is added to the relationship between the MNO and the customer as a so-called third party whose services must be charged in any manner. Typically, B2C value-added services are charged by the MNO that has an existing billing relationship with customers. We refer to this as third-party billing. The MNO performs data transmission and also acts as a payment service provider for the content provider. Therefore he charges a higher fee to the customer which includes the value of the content as well as its transmission.

Charging Concepts for Third-Party Billing

Basically, there are three basic *charging concepts* for third-party billing: sponsoring, charging by premium fee, and charging by fixed price (Turowski & Pousttchi, 2004). In this context the *sponsoring* concept plays a special role. Services are free of charge for customers, because they are provided at the mobile content provider's expense. The mobile content provider may even pay the MNO for data transmission, offering not only the content, but even the access to it completely for free to the customer. Sponsoring models are not common up to now. A typical example for a completely free service would be a car manufacturer who offers a video download with a commercial for its new car model. Sponsoring models will also become relevant if business models financed by advertisement are applied or if the free service aims at the usage of further service offer with costs.

For the two other charging concepts, customers will pay the MNO by the mobile phone bill. Both are based on the principle of revenue sharing between MNO and mobile content provider. From the perspective of most MNOs, the preferred solution is *premium fee charging,* meaning that customers pay a data volume fee for transmission and additionally a so-called premium fee for the value of the content or service. The MNO as payee gets the data volume fee and transfers the premium rate to the mobile content provider after deducting a compensation for his charging costs. The simplest example for charging by premium fee is charging services by premium SMS or premium MMS. Customers pay for sending (or, as common in some markets, also for receiving) the SMS a higher fee consisting of the SMS basic price and a premium rate. The latter underlies a revenue sharing. In case of singly ordered SMS/MMS services, charging by premium fee is simple and broadly accepted, although other services demand a more complex application.

However, apart from SMS, the existence of a data volume fee causes acceptance problems on the customers' side. On the one hand customers do not benefit from the size of transmitted volume; on the other hand customers with low affinity to technology are not able to imagine how much for example one kilobyte of data is. Figure 1 explains the premium fee charging concept; the delineation follows the e^3-value method of Gordijn, Akkermans, and van Vliet (2000) and shows the benefits as value flows between the parties of the stakeholder network which are therefore represented by arrows.

From the perspective of most consumers, the preferred solution is *fixed price charging.* Customers pay a fixed price for the usage of the service. This revenue is shared between MNO and mobile content provider according to an agreed ratio. The problem with this solution lies in the fact that "real" revenue sharing is necessary (whereas beforehand, transmission and content have been paid separately). Hence, the ratio must include the transmission costs of the service. This concept is called "airtime revenue sharing" and seen as extremely unpleasant by the vast majority of MNOs, as they hope to compensate declining voice revenues with this and thus defend their margins strongly. Figure 2 illustrates the fixed price charging concept.

FUTURE TRENDS

Mobile payment is highly attractive to consumers. Outside mobile commerce we have to distinguish two different goals: in markets with an established banking infrastructure such as Europe, the United States, or Australia, this concerns a new channel to existing payment methods and is primarily a

Figure 1. Premium fee charging

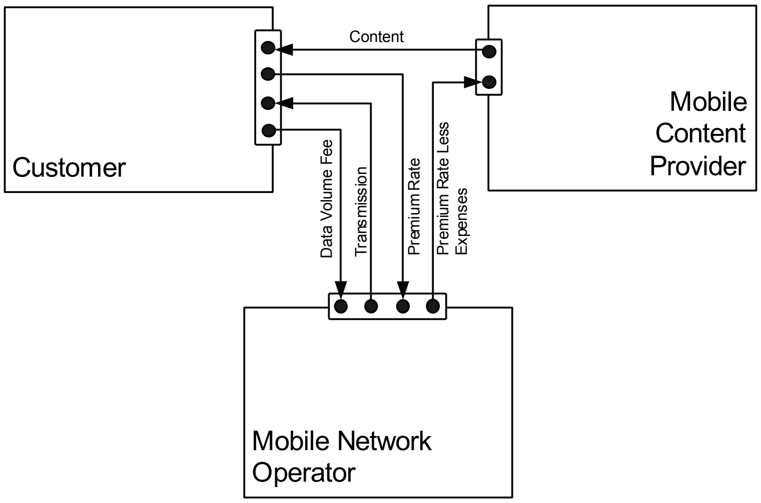
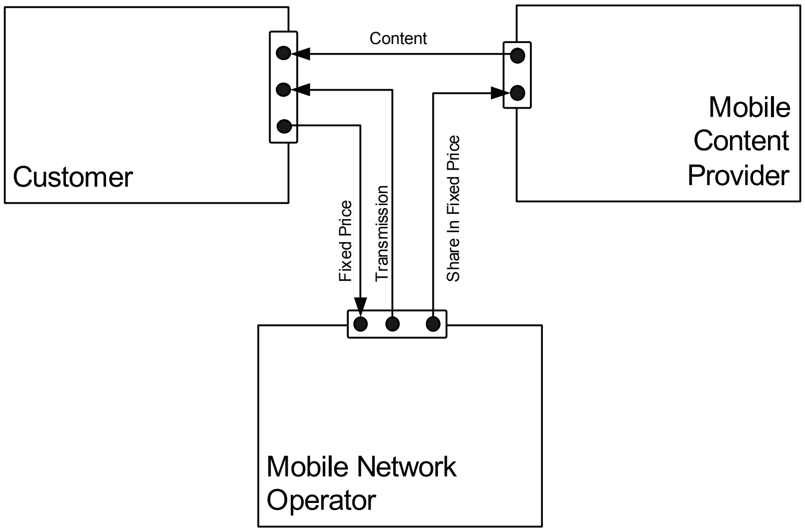

Content
Mobile
Content
Provider
Customer
Data Volume Fee
Transmission
Premium Rate
Premium Rate Less
Expenses
Mobile Network
Operator

Figure 2. Fixed price charging


Content
Mobile
Content
Provider
Customer
Fixed Price
Transmission
Share In Fixed Price
Mobile Network
Operator

question of convenience. In emerging markets such as many countries in Africa or South America, mobile payment is more likely to become a new payment method. The aim for the latter, under the keyword "banking the unbanked," is to enable millions of people who do not own a bank account but a mobile phone to replace cash transactions. Payment service providers with a focus outside of mobile commerce are often bank-driven.

Mobile payment inside mobile commerce is closely related to the core business of the MNOs who, in most markets, control the access to and charging for mobile content. Often they do so without caring about customer preferences, or they even result in obstructing the market for mobile content. On the other hand they come more and more under pressure by market fragmentation through MVNO models and declining voice margins. Current trends show a de-globalization of the MNO business, for instance in the failure of i-mode outside of Japan, in the failure of Vodafone in the Japanese market, and in the failure of a European MNO to play a major role in the U.S. market. The consequences of this de-globalization for payment and charging strategies of the MNO promise to be interesting.

CONCLUSION

Mobile payment is seen as an important issue since the mid-1990s, first as an m-commerce application to complete payments in the "real world" and secondly in order to charge for third-party mobile services. Both of these are still promising, but not yet successful in most markets.

In order to enable uptake of mobile services, MNOs have to offer payment methods that follow customer preferences, especially fixed-price charging, and that provide incentives to the owners of valuable content, especially airtime revenue sharing.

De-globalization of MNO business and increased pressure on MNOs may lead to easier and more open approaches to mobile payment on national markets, including banks, merchants, and content providers in these markets.

REFERENCES

Coursaris, C., & Hassanein, K. (2002). Understanding m-commerce—A consumer centric model. *Quarterly Journal of Electronic Commerce, 3*(2), 247-271.

Dahlberg, T., Mallat, N., & Öörni, A. (2003). Consumer acceptance of mobile payment solutions—Ease of use, usefulness and trust. In G. M. Giaglis, H. Werthner, V. Tschammer, & K. Froeschl (Eds.), *Proceedings of the International Conference on Mobile Business* (pp. 211-218). Vienna, Austria.

Eisenmann, M., Linck, K., & Pousttchi, K. (2004). Use case scenarios for mobile payment procedures—Results of the study MP2. In K. Pousttchi, & K. Turowski (Eds.), *Proceedings of the Workshop on Mobile Commerce* (pp. 50-62). Augsburg, Germany.

Gordijn, J., Akkermans, J. M., & van Vliet, J. C. (2000). What's in an electronic business model? In R. Dieng & O. Corby (Eds.), *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management* (pp. 257-273). Juan-les-Pins, France.

Henkel, J. (2002). Mobile payment. In G. Silberer, J. Wohlfahrt, & T. Wilhelm (Eds.), *Mobile commerce—Basics, business models and success factors* (pp. 327-351). Wiesbaden: Gabler.

Khodawandi, D., Pousttchi, K., & Wiedemann, D. G. (2003). Acceptance of mobile payment procedures in Germany. In K. Pousttchi & K. Turowski (Eds.), *Proceedings of the Workshop on Mobile Commerce* (pp. 42-57). Augsburg, Germany.

Kieser, M. (2001). Mobile payment—Comparison of electronic payment systems. In A. Meier (Ed.), *Mobile commerce* (HMD Vol. 220, pp. 27-36). Heidelberg: Dpunkt.

Kreyer, N., Pousttchi, K., & Turowski, K. (2002). Standardized payment procedures as key enabling factor for mobile commerce. In K. Bauknecht, G. Quirchmayr, & A. M. Tjoa (Eds.), *Proceedings of the International Conference on EC-Web* (pp. 400-409). Aix-en-Provence, France.

Krueger, M., Leibold, K., & Smasal, D. (2006). *Online payment methods from the viewpoint of customers—Results of the study IZV8.* University of Karlsruhe, Germany.

Mallat, N. (2004). Theoretical constructs of mobile payment adoption. *Proceedings of the Information Systems Research Seminar in Scandinavia (IRIS)*. Falkenberg, Sweden.

MobilMedia. (2004). *MobilMedia barometer: Second wave: M-payment.* Representative survey with 567 informants conducted by the agency Brodeur on behalf of the German Ministry of Economics initiative MobilMedia under academic supervision of K. Pousttchi, Berlin.

Natsuno, T. (2003). *The i-mode wireless ecosystem.* New York: John Wiley & Sons.

Pousttchi, K. (2003). Conditions for acceptance and usage of mobile payment procedures. In G. M. Giaglis, H. Werthner, V. Tschammer, & K. Froeschl (Eds.), *Proceedings of the International Conference on Mobile Business (pp. 201-210)*. Vienna, Austria.

Pousttchi, K., Selk, B., & Turowski, K. (2002). Acceptance criteria for mobile payment. In F. Hampe & G. Schwabe (Eds.), *Proceedings of the Conference on Mobile and Collaborative Business, Multikonferenz Wirtschaftsinformatik* (pp. 51-67). Nuremberg, Germany.

Pousttchi, K., & Wiedemann, D. G. (2006). Charging mobile services in the mobile-payment-reference model. In T. Lammer (Ed.), *Handbook e-money, e-payment & m-payment* (pp. 363-378). Heidelberg: Physica-Verlag.

Taga, K., & Karlsson, J. (2004). *Arthur D. Little global m-payment report.* Vienna: Arthur. D. Little Austria GmbH.

Turowski, K., & Pousttchi, K. (2004). *Mobile commerce* (1st ed.). Heidelberg: Springer.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*(3), 185-198.

Wirtz, B. (2001). *Electronic business* (2nd ed.). Wiesbaden: Gabler.

Zobel, J. (2001). *Mobile business and m-commerce—Conquering future markets* (1st ed.). Munich: Hanser.

KEY TERMS

Direct Offer by the Mobile Content Provider: Offer model for mobile services in which mobile content providers have a direct customer relationship and generate direct added value by the service offer.

Fixed Price Charging: Charging concept in which customers pay a fixed price for the mobile services which is shared between the MNO and the mobile content provider.

Mobile Billing: Billing of telecommunication services by an MNO within the scope of an existing billing relationship.

Mobile Commerce: Any kind of business transaction, on the condition that at least one side uses mobile communication techniques.

Mobile Payment: Type of payment transaction processing in the course of which—within an electronic procedure—at least) the payer employs mobile communication techniques in conjunction with mobile devices for initiation, authorization, or realization of payment.

Offer by the MNO: Offer model for mobile services in which MNOs produce mobile services or buy them from a mobile content provider, and the offer appears as MNO and mobile content provider simultaneously.

Premium Fee Charging: Charging concept in which customers pay a data volume fee for the transmitted data volume and additionally a premium rate for the value-added service.

Sponsoring: Charging concept in which services are free for customers, because they are provided at the mobile content provider's expense.

Standard Type Phone Bill: The mobile payment standard type is characterized by an MNO as mobile payment service provider and a mobile phone bill as settlement method.

Mobile Phone Gambling

Mark Griffiths
Nottingham Trent University, UK

INTRODUCTION

It is often claimed by marketeers that online gambling (i.e., the combining of gambling and the Internet into one convenient package) makes commercial sense. Gambling looks like it might make another step towards convenience with the advent of mobile phone gambling. This is gambling on the move, whenever, wherever, with the wireless world of mobile gambling. Since it is somewhat unnatural to always be near a computer, it could be argued that wireless mobile phones are the ideal medium for gambling. Whenever gamblers have a few minutes to spare (at the airport, commuting to work, waiting in a queue, etc.), they can occupy themselves by gambling (Griffiths, 2005a).

BACKGROUND AND BRIEF OVERVIEW

The mobile phone industry has grown rapidly over the last few years. It is predicted that by the end of 2005, the number of international mobile phone users will pass the two billion mark. The latest research by Mintel (2005a) highlights that mobile phone revenues from mobile gambling and gaming is increasingly rapidly. In 2004, mobile gaming revenue reached $200 million. According to the Mintel report, by 2009, mobile gambling is set to generate $3 billion in the United States alone. Despite the huge figure, mobile gambling will only likely account for around 1.5% of mobile industry revenues. It will also be a small part of the overall market as Mintel predicts that the U.S. casino gambling market will generate revenues of almost $71 billion by 2009 (compared to the $48.3 billion generated in 2004).

In the UK, mobile phone gambling has also increased dramatically. Mintel (2005b) reported that the number of betting pages downloaded by the end of 2005 was expected to approach three million, up 367% on 2004. Mobile phone users in the UK are set to spend £740m on phone downloads by the end of 2005. This is 18 times the £40m spent in 2002. Ring tones for phones account for approximately one-third of all mobile downloads. Arcade-style games (26%), screensavers and wallpaper (13%), and music (8%) are all popular. However, the biggest growth area has been in gambling, which now accounts for 9% of all mobile phone downloads in the UK.

These predictions also seemed to be backed up by Juniper Research (2005) who predicted mobile gambling revenues will total about $19.3 billion worldwide by 2009, with lotteries accounting for about $7.9 billion, sports betting bringing in $6.9 billion, and casino-style gambling contributing $4.5 billion. Juniper predicts that lotteries will make most money for mobile gambling operators because governments are generally less censorious about lotteries than other forms of gambling. They are also easy to play and relatively low cost compared to other types of gambling (Griffiths & Wood, 2001). This means that mobile lotteries are likely to become established fairly quickly in a greater number of markets. Given the ubiquity of lotteries worldwide, it only requires a very small percentage of players to buy their tickets via their mobile phone for the resulting global dollar revenues to run into the billions. Juniper also claims that the growth in the UK National Lottery is almost wholly attributable to mobile betting.

Conventional wisdom says that two things have the power to drive any new consumer technology—pornography and gambling (Griffiths, 2001, 2004a). These activities helped satellite and cable television, video, and the Internet. It has been claimed in the media that Internet gambling and adult (pornography) sites are about the only e-businesses easily succeeding, as they provide adult entertainment in a convenient and guilt-free environment. The wireless world of the mobile phone may not be too different. So will gambling compete with pornography for dominance of mobile commerce? Along with pornography, gambling sites are one enterprise that should have little trouble reaching profitability—especially if this is combined with sports events. Sports are huge on the Internet. There are thousands of communities on the Internet built around sports teams or leagues, and even more "unofficial" team sites set up by fans. The most successful of those communities will look to "mobilize" and then "monetize."

To some extent, the majority of gamblers are risk-takers to begin with (Griffiths, 2004b). Therefore, they may be less cautious with new forms of technology. Third-generation (3G) mobile phones are ideal for bet placing, and gamblers will be able to check on their bets and place new ones. Furthermore, it is anonymous and can provide immediate gratification, anytime, anywhere. Anonymity and secrecy may be potential benefits of mobile gambling, as for a lot of people there is still stigma attached to gambling in places like betting shops and casinos (Griffiths & Parke, 2002). Mobile gambling is also well suited to personal (i.e., one-to-one) gambling, where users bet against each other rather than bookies. Online betting exchanges (e.g., betfair.com)

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

are prime examples of where people bet on anything and everything with each other (Griffiths, 2005b).

So what types of gambling will work best on mobile phones? Internet gambling lends itself most naturally to "casino-style" games like slot machines, blackjack, roulette, poker, and so forth. These games require more in the form of graphics, sound, and interactivity. These types of gambling are not ideal for mobile devices but have now been introduced on 3G devices. It is unlikely that mobile phone graphics and technology can compete with Internet Web browsers (although the technology is improving all the time). Intuitively, mobile phone gambling is best suited for sports and event betting. With mobile phone betting, all that is required is real-time access to data about the event to be bet on (e.g., a horse race, a football match), and the ability to make a bet in a timely fashion.

These basic requirements are easily provided by the current generation of mobile phones and the appropriate software. At present, it looks as though mobile phones' biggest influence will be on sports betting. The placing of the bet is not the driving motivation in event wagering. Since being the spectator is what sports fans are really interested in, the sports gambler does not need fulfillment from the process of gambling. People betting on sports will use mobile phones because they are easy, convenient, and take no time to boot up. Once they have their sports book registered as a bookmark on their phone, they can access it and place a bet within minutes.

FUTURE TRENDS

However, the situation will change over time. The new generation of mobile phones already has the capability to play typical "casino-style" games like blackjack, poker, and slots. The limiting aspects of the technological and protocol demands of mobile gambling (graphics, sound, and displays on mobile and personal digital assistant devices) are largely being resolved through technological advance.

These advances will allow punters to watch sporting events live on their phones while wagering in real time. Consider the following scenario. A betting service that knows where you are and/or what you are doing has the capacity to suggest something context related to the mobile user to bet on. For instance, if the mobile phone user bought a ticket for a soccer match using an electronic service, this service may share this information with a betting company. If in that match the referee gives a penalty for one team, a person's mobile phone could ring and give the user an opportunity (on screen) to bet whether or not the penalty will be scored. On this type of service, the mobile phone user will only have to decide if he or she wants to bet, and if so, the amount of money. Two clicks and the bet will be placed. Context, timeliness, simplicity, and above all user involvement look like enough also to convince people who never entered a bet shop.

The scenario described is not as far-fetched as it would seem. Manchester United soccer club has transformed itself into a powerful media company. It has launched its own digital TV channel, signed up a host of big-name technology partners (including Vodaphone, Sun, Lotus, and Informix), and started an ISP service. Its partnership with Vodaphone is perhaps a sign of the shape of things to come. In addition to sponsoring the club's kit, Vodafone will also get the chance to develop co-branded mobile services with the club. This will offer users access to content similar to the company's Web site (receiving real-time scores and team news via SMS). What Vodafone is heading towards is the ultimate goal—live video of matches, straight to mobiles, anywhere in the world.

While watching matches, users will be able to view statistics, player biographies, and order merchandise. So what does all this have to do with gambling? Mobility will facilitate an increase in "personalized" gambling, for example, the types of service offered by Eurobet's Match service, where bettors gamble against each other, rather than the house.

Gambling will become part of the match day experience. A typical scenario might involve a £10 bet with a friend on a weekend football match. The gambler can text the friend via SMS and log on to the betting service to make a bet. If the friend accepts, the gambler has the chance to win (or lose). Football clubs will get a share of the profits from the service. Clubs are keen to get fans using branded mobile devices where they can simply hit a "bet" button and place a wager with the club's mobile phone partner.

The penetration of wireless gambling will mostly be contingent upon the market penetration of wireless Web users in general. The mobile phone market is already large in many parts of the world. Juniper predicts that by 2009, mobile gambling revenues will be concentrated in Europe (37%) and the Asia-Pacific region (39%). They predict that North America will produce only 15% of global revenues because of government and societal opposition to wireless gambling. If these numbers are combined with the popularity of gambling, it could be speculated that there is the basis for a very profitable enterprise.

As with all new forms of technological gambling, ease of use is paramount to success. In the early days of WAP phones, programming the phone to use the protocol was very difficult. However, mobile phones are becoming more user friendly. Pricing structures are also important. Internet access and mobile phone use that is paid for by the minute produces very different customer behavior to those that have one-off payment fees (e.g., unlimited use and access for a monthly rental fee). The latter payment structure would appear to facilitate leisure use, as consumers would not be worried that for every extra minute they are online, they are increasing the size of their bills.

SOCIAL IMPACT: SOME CONCERNS

As with all new forms of technology—especially when used for gambling—there are some areas of potential concern. These are briefly outlined below (adapted from Griffiths, 2004c).

Access and Convenience

As already (implicitly) mentioned, mobile gambling's greatest advantage (even over the Internet) is its accessibility. Gamblers with mobile phones are no longer bound to computers and Internet access. The only thing that separates a mobile phone gambler from his or her favorite games is network coverage. Regardless of location, anyone can bet on their favorite sporting event or play their favorite casino games via mobile phone. Mobile phone gambling eliminates geographical borders, travel, and queuing up to bet. With mobile gambling, anyone can (theoretically) gamble 24 hours a day, seven days a week from the gambler's preferred location. It could also be argued that mobile phones make "impulse betting" easier. Like Internet gambling, it is also another example of convenience gambling (Griffiths, 2003a, 2005a). Given that prevalence of behaviors is strongly correlated with increased access to the activity, it is not surprising that the development of regular mobile use is increasing across the population. Increased accessibility may also lead to increased problems. Research into other socially acceptable but potentially addictive behaviors (drinking alcohol) has demonstrated that increased accessibility leads to increased uptake (i.e., regular use) and that this eventually leads to an increase in problems (although the increase may not be proportional) (Griffiths, 2003b).

Targeting the Low Earners

A popular claim by anti-gambling campaigners is that the low-limit bets and the relatively low payouts on many types of mobile phone gambling will attract mostly lower-class gamblers. However, most low-income workers are unlikely to own a brand-new high-tech PDA or mobile phone that incorporates such technology and services (although as prices continue to decrease, this may change). According to the gaming industry, mobile phone betting requires a very high standard of security and reliability that can only be reached with the newest (and most expensive) mobile phones. Regardless of disposable income, mobile phone gambling providers must (at the very least) allow gamblers to set pre-determined spending limits. This will help the gambler to avoid chasing losses (a known risk factor that facilitates problem gambling).

Youth Gambling

It has also been claimed that promoters of mobile phone gambling think it will attract a younger breed of gambler. In fact some companies are deliberately targeting the under-16 market with mobile phones specially designed for them, although they are not targeting gambling per se. This is something that needs to be monitored (Smeaton & Griffiths, 2004). Mobile phones that do not implement a user ID program will be very hard to trace and check—in particular when it comes to under-age customers trying to place a bet. However, the industry claims one way to tackle underage use is through a pre-paid card system. Pre-pay systems are bound by the same security and accounting best practices currently employed within the casinos today. This may potentially minimize problem gambling and prevents access to minors since distribution is controlled by the operator.

As we can see, potential social impact always follows new developing markets. Mobile phone gambling is clearly an area that needs in-depth monitoring of the psychosocial impact over the next few years.

CONCLUSION

It is clear that mobile phone gambling is still a relatively untapped area, and the functional capabilities of mobile phones are getting better all the time. Cell phones are rapidly growing in their functional capabilities. Mobile gambling is available on most of the mobile phones that are powered by Windows Mobile, Symbian OS, RIM including Java, and browser-based phones. There are now Internet sites that allow mobile phones to download casino-style games to the gambler's phone, allowing real money betting from anywhere they can get a phone signal. As the new generation of mobile phones accept Java programming, the high-end graphic display can be used to deliver live video feeds for the various casino games.

It appears that sophisticated mobile phone technology is increasingly able to integrate within our culture. This will have implications for the social impact and will need monitoring. The research by both Mintel and Juniper raises the possibility that almost unlimited access to mobile phone gambling will lead to more problem gambling. Like Internet gambling, mobile phone gambling has completely changed the way people think about betting. Mobile phones provide the convenience of making bets or gambling from wherever the person is. On paper, this all sounds relatively simple and is set to get even easier. Many gaming industry observers are claiming that in the not too distant future, people will not go to sporting events like horse races or football anymore. They will simply watch the sport on television and place bets via their mobile phones.

REFERENCES

Griffiths, M. D. (1999). Gambling technologies: Prospects for problem gambling. *Journal of Gambling Studies, 15,* 265-283.

Griffiths, M. D. (2001). Sex on the Internet: Observations and implications for sex addiction. *Journal of Sex Research, 38,* 333-342.

Griffiths, M. D. (2003a). Internet gambling: Issues, concerns and recommendations. *CyberPsychology and Behavior, 6,* 557-568.

Griffiths, M. D. (2003b). Adolescent gambling: Risk factors and implications for prevention, intervention, and treatment. In D. Romer (Ed.), *Reducing adolescent risk: Toward an integrated approach* (pp. 223-238). London: Sage.

Griffiths, M. D. (2004a). Mobile phone gambling: Preparing for take off. *World Online Gambling Law Report, 8*(3), 6-7.

Griffiths, M. D. (2004b). Betting your life on it: Problem gambling has clear health related consequences. *British Medical Journal, 329,* 1055-1056.

Griffiths, M. D. (2004c). Interactive television gambling: Should we be concerned? *World Online Gambling Law Report, 3*(3), 11-12.

Griffiths, M. D. (2005a). Remote gambling: Psychosocial aspects. In *Proceedings of Remote Gambling* (Westminster E-Forum Seminar Series) (pp. 11-20). London: Westminster Forum Projects Ltd.

Griffiths, M.D. (2005b). Online betting exchanges: A brief overview. *Youth Gambling International, 5*(2), 1-2.

Griffiths, M. D., & Parke, J. (2002). The social impact of Internet gambling. *Social Science Computer Review, 20,* 312-320.

Griffiths, M. D., & Wood, R. T. A. (2001). The psychology of lottery gambling. *International Gambling Studies, 1,* 27-44.

Juniper Research. (2005). *Betting on mobile gambling.* Retrieved October 11, 2005, from http://www.spin3.com/in070305mobilegambling.php

Mintel. (2005a). *Mobile gambling expected to sweep the U.S.* Retrieved October 3, 2005, from http://www.spin3.com/sweep.php

Mintel. (2005b). *Mobile phone gambling on the increase.* Retrieved October 3, 2005, from http://www.onlinecasinonews.com/ocnv2_1/article/article.asp?id=8684

Smeaton, M., & Griffiths, M. D. (2004). Internet gambling and social responsibility: An exploratory study. *CyberPsychology and Behavior, 7,* 49-57.

KEY TERMS

Convenience Gambling: Remote forms of gambling such as Internet gambling, interactive television gambling, and mobile phone gambling. Convenience gambling allows the gambler to gamble when he or she wants to without leaving the comfort of home and/or the workplace.

Java: A feature that allows the device to run specially written applications. Java applications can provide specific functions such as games, or they can be custom-written corporate applications. Some phones allow the user to download new applications directly from Internet, while others require a data cable to transfer the applications from a PC.

Mobile Gambling: One of several mobile (cellular) phone activities that are similar to online (Internet) gambling and offer the chance for players to win money on a range of different gambling activities.

Mobile Gaming: Video games played on mobile (cellular) phones. There is no opportunity to win money on these types of game.

Mobile Phone: A wireless telephone that sends and receives messages using radiofrequency energy in the 800-900 megahertz portion of the radiofrequency (RF) spectrum. Also called a cell phone.

Short Message Service (SMS): A service that enables subscribers to send short text messages (usually about 160 characters) to and from mobile phones.

Third Generation (3G): Analog cellular phones were the first generation. Digital marked the second generation. 3G is loosely defined, but generally includes high data speeds, always-on data access, and greater voice capacity. The high data speeds are possibly the most prominent feature and certainly the most hyped. They enable such advanced features as live, streaming video.

Mobile Phone Privacy Issues

Călin Gurău
Montpellier Business School, France

INTRODUCTION

Mobile phones have become a normal feature of our social life. People use them in institutions, on the street, in buses or trains, in restaurants, and even while driving, although in many countries this is a forbidden practice. The ringing of a mobile phone has become a familiar sound, and people have grown accustomed to witnessing loud-voice telephone conversations. However, not everybody is happy about this additional noise, which can be considered as an invasion of the personal privacy (Monk, Carroll, Parker, & Blythe, 2004).

On the other hand, the new generations of mobile phones started to incorporate advanced computing and communication facilities, such as location-tracking or position-aware applications, which can be used by mobile phone companies, relatives and friends, or third parties to identify the specific location of a mobile phone user (Broache, 2006). With the introduction of SMS and mobile Web browsers, the phenomenon of spam has also infected mobile phones. A new generation of hackers started to attack mobile devices, creating specific viruses. Recently, a company called Vervata announced the development of the first spyware for mobile phones—FlexiSPY—which is "absolutely undetectable by the user," and which can be used to monitor the SMS messages that are sent and received, or to record the duration and the history of these calls. The data captured is then sent to Vervata's servers, being accessible to customers via a special Web site (Evers, 2006). In some cases, governments have considered the possibility to register and analyze mobile phone conversations in order to track down and prevent harmful activities such as terrorism (Lettice, 2003; Richtel, 2005). All these issues raise important concerns regarding the personal privacy of the mobile phone users.

This article presents an overview of the main privacy problems raised by mobile phones, both for their users and for society in general, and analyzes a series of solutions for reducing their negative impact. After discussing the evolution of private vs. public space in the last decade, the article presents the main privacy concerns related with mobile phone usage. A series of data collected both from mobile phone owners, as well as from people without mobile phones, is analyzed, providing a basis for comparing conflicting opinions related to privacy issues. The article concludes with a series of solutions that already exist or may be available in the near future, and which can solve or reduce mobile phone privacy problems.

BACKGROUND

The use of mobile phones is becoming truly universal. The average mobile phone ownership in Europe is above 55%, with Spain, Norway, Iceland, and the Czech Republic having more than 90% cell phone coverage in the population, and Luxembourg, Taiwan, Italy, and Hong Kong reaching 100% (Kristoffersen, 2005; Plant, 2001).

The use of mobile phones has determined a series of contradictory changes in the social and work environment (Townsend, 2000). The importance of investigating these changes stems from the fact that the mobile phone represents only the first step toward the introduction of ubiquitous computing, a trend demonstrated by the redefinition of mobile phones from 'communication devices' to 'mobile platforms for computing and networking applications'.

Research conducted into the social impact of mobile phones has identified their multiple functionality—mobile phones are perceived as communication devices, but also as identity signifiers and fashion symbols. People can use them not only to communicate with their friends or relatives, but also to express their identity by choosing the color of phone or a specific ring tone. The mobile phone is considered by many people as a part of their personal identity (Hulme, & Peters, 2001).

Social research has emphasized that mobile phones force people to combine different aspects of their life, as for example when a person receives a call with a subject that is totally at odds with their physical situation. In such a situation the mobile phone user looks embarrassed and ill at ease, and often she or he tries to find an isolated place in which she or he can talk openly (Geser, 2004). Other people become experts in managing and dealing with the two conflicting situations simultaneously, for example riding a bicycle and taking at the same a mobile call (Plant, 2001). From this perspective, mobile phone technology has blurred the boundaries between private and public space, creating situations in which the two co-exist and blend their characteristics (Geser, 2004).

The use of mobile phones changes the social relations among people, and the positioning of a person both towards

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

close friends and relatives, and toward other citizens. It is easier to contact a friend who has a mobile phone anytime of the day, but on the other hand, the mobile phone might increase the isolation of an individual in a crowd, because s/he will be less inclined to ask information or help from a passerby. As a consequence, mobile phone users might lose some essential social skills, such as their capacity to communicate with strangers (Fortunati, 2002).

Despite the increased interest about the privacy problems raised by mobile phones, most of the existing studies attempt to investigate only one specific privacy issue (Hulme & Peters, 2001; Kindberg, Spasojevic, Fleck, & Sellen, 2004; Minch, 2004), without attempting to provide a clear general overview of the perceived privacy threats. This study attempts to fill this knowledge gap, presenting the results of an empirical research about the perception of privacy problems in two groups of respondents—owners and non-owners of mobile phones.

PRIVACY RELATED TO MOBILE PHONE USAGE

Privacy was constantly considered a central issue in relation to the use of information and communication technologies. However, its meaning has evolved in direct relation to the social transformations determined by the evolution of technology. Palen and Dourish (2003) have analyzed the issue of privacy within a networked environment, using the theoretical framework developed by Altman. Altman (1975, 1977) argues that privacy is not a fixed and immutable concept, but represents a selective control of access to the personal self, regulated through a process of dialectic and dynamic boundary regulation. For Altman, privacy does not simply mean avoiding information disclosure, but rather a selective disclosure of personal information which permits reconciliation of the desire for a private life with the maintenance of an accepted social persona.

The realization and maintenance of this equilibrium between private and social spheres are increasingly complex in a networked society, in which the very existence of a person is linked to some form of information sharing. The extreme case of total privacy control will imply an asocial life, which rejects not only the disadvantages but also the advantages of communication and social involvement.

The case of mobile phones is rather symptomatic for this situation: the use of mobile phones provides clear personal and social advantages, together with a series of threats and problems concerning the management of personal information. The solutions to these privacy issues should be identified and applied intelligently in order to eliminate or reduce the problem, but also to maintain the advantages offered by mobile communication and computing.

The first logical step in solving mobile phone privacy problems is to identify the range and intensity of these issues. The studies investigating these dimensions are few and dispersed, usually presenting a theoretical framework for mobile phone privacy issues, which is often unsupported by empirical evidence and data analysis.

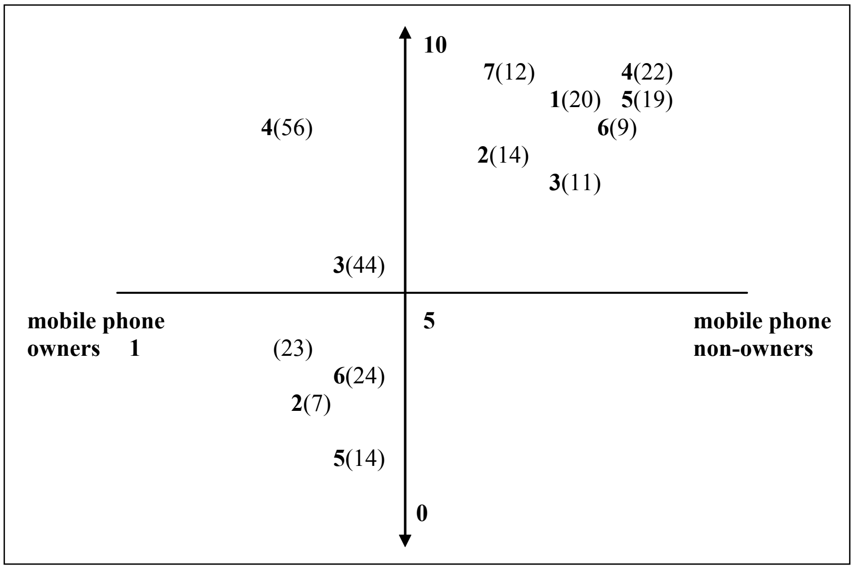
In relation to the use of mobile phones, previous studies' threats have identified the following privacy issues for mobile phone owners and/or for the people from their entourage: (1) location tracking (Minch, 2004); (2) mobile phone conversations that are heard by other people (Geser, 2004); (3) data interception by governmental authorities (Richtel, 2005); (4) data interceptions by third parties—spywares (Evers, 2006); (5) the noise of ring tones and mobile phones conversations (Geser, 2004; Plant, 2001); and (6) the use of mobile phones to take pictures (McLeod, 2003).

In order to identify the perceptions of people regarding these privacy issues, a short face-to-face interview was applied to two groups of people, one comprising mobile phone owners (56 people) and another formed by people who do not own a mobile phone (22 people). The interviews took place between April and June 2005 in Montpellier, France, and lasted about 15 minutes. The respondents were asked to rate the importance they attach to various mobile phone-related privacy issues on a scale from 0 (no importance) to 10 (very high importance). The questions focused on the six main privacy issues already identified in the literature, but at the end an open question was added, asking the respondents to specify any other perceived privacy threat related to mobile phones.

The positioning map reproduced in Figure 1 offers an overview of the number of people concerned about a certain privacy issue and of the intensity of their concern, allowing also an easy comparison between mobile phone owners and non-owners. The numbers before parentheses represent the code of the privacy issue being identified and considered by respondents, while the numbers within parentheses represent the number of people from each of the two groups that have indicated a specific privacy issue as important. Finally, the position of each group of symbols indicates the mean importance of each privacy issue calculated on the basis of the importance levels provided by the respondents (from the two groups) who considered the privacy issue as being relevant for them.

Some of the respondents from the non-owners' group have indicated an additional privacy concern, which, in some respects, can be considered as a derivation of location tracking privacy threats, namely availability. These respondents indicated that the ownership and use of a mobile phone directly indicate the availability to be contacted by various people and organizations, contacts that can impact on the structure and content of your private life, and which, in some cases, can restrict your personal privacy. This concern is a good illustration of the fear of conflict between the private

Figure 1. The spatial representation of various mobile phone-related privacy issues, and of their importance for owners and non-owners of mobile phones



and the public spheres (one respondent gave the example of receiving a mobile phone call while seeing a good movie at a cinema), which forces the user to re-adapt quickly to the requirements of the 'other sphere' and neglect the actions/circumstances in which s/he is involved at that moment. The fact that a number of non-owners have indicated this as a fairly important issue may indicate this as a potential reason in their decision to not own a mobile phone device.

It is also interesting to note that the distribution of answers provided by the members of the two groups is clearly different in terms of the perceived importance of various privacy issues. Overall, the non-owners of mobile phones perceive the privacy-related problems as more important that the group of mobile phone owners. The only exception is the issue of 'data interception by third parties using spywares', in which both groups give a high score (8.1 by owners of mobile phones, and 9.2 by the non-owners). The more negative perception of the non-owners group can represent a certain bias in their judgment of the privacy risks of using a mobile phone—a bias which may have prevented them from using a mobile device.

FUTURE TRENDS

The future development of information and communication technology and applications clearly indicates that the privacy-related issues will continue to represent an object of debate. On the other hand, the concept is expected to evolve, in line with the social and technological changes in the society and in people's lifestyles.

The future evolution of mobile telephony and the privacy issues related to the use of mobile phones can be analyzed from three inter-related perspectives: technological, social, and legal.

From a technological point of view, the miniaturization of mobile devices and the integration of multiple functionalities and applications will probably continue. Mobile phones will represent the basis for the new concept of ubiquitous computing. The communication will take place not only between people, but also between humans and intelligent software applications, or through digital intelligent agents that will be stored in the computing devices dispersed in the environment. The main problem related with this technological trend is the need to improve computing and information transfer security. With personal communication flowing among various computing devices, the risk of data interception will be multiplied; therefore better firewalls and encryption programs will have to be implemented. On the other hand, as voice recognition applications will be gradually introduced, biometric security tests based on voice profile will complement the present system of personal passwords. Personal intelligent agents will have the capability to dynamically negotiate the privacy levels offered by various software applications, using the elements provided by their users.

The social landscape of mobile communication will be influenced by the principles of the networked society. The non-owners of mobile phones will limit their capacity of interacting with various computing devices that provide useful services and information. They will be able to preserve their personal privacy, but only with the cost of being disconnected from the digital social network. On the other hand, the present problem of having the personal or social space invaded by other people who use their mobile phones will continue, although new solutions can be found to alleviate the discomfort (e.g., the replacement of ring tones with more discrete methods to indicate an incoming call).

Some sociologists (Geser, 2004; Plant, 2001) believe that the increased use of mobile phones will accelerate the dissolution of the traditional family. Using mobile phones, the members of a family will be able to communicate easier with their closer or distant relatives, but on the other hand, they will develop an independent communication network, linking with friends and participating to spatially dispersed communities of interest. Nowadays children already use mobile phones as an expression of their communicational independence, because the personalized number allows them to reduce the transparency of their communication networks to parents.

Finally, at a legal level, the use of mobile phones and the privacy problems developed in relation to this use often determines the introduction of new legislation, which has a large potential impact on all forms of communication. Following the introduction of digital cameras in mobile devices, and some serious offences against privacy, Victoria and Western Australia introduced new surveillance devices legislation that restricts how photographs of private activities may be taken and/or used. Many voices have even asked for bans on the use of phone-cameras, but any wrong initiative can have an impact not only on the use of mobile phones, but also on the way information is transmitted and used. For example, the way in which the press is taking and using pictures of people and events can be fundamentally transformed by such legal initiatives (McLeod, 2003).

In many countries, restrictions have been imposed for the use of mobile phones in specific places. Trains in Britain, Japan, Switzerland, and the U.S. now have quiet carriages; and restaurants in cities as diverse as Cairo and Chicago have introduced 'no-mobile' policies, or 'mobile-free' zones in order to protect the privacy and personal space of their customers (Plant, 2001).

In other situations, governments have attempted to introduce legislation that permits the policy or the security forces to intercept and analyze mobile phone conversations in order to prevent criminal or terrorist acts (Perera, 2001). Although the legislative bodies have sometimes approved such legislative initiative, the measures have been received by the population with an increased concern regarding the limitations of personal privacy (Richtel, 2005).

The regulation of mobile phone use will continue to evolve, illustrating the conflict among the free use of mobile phone technology, the requirements for personal privacy, and the governmental surveillance for crime prevention purposes.

CONCLUSION

This article attempted to present a general overview of the main privacy issues raised by the increased used of mobile phones in modern society. Both the examples and the empirical study provided in this article indicate that mobile phone privacy-related problems are complex and multifaceted. The large-scale introduction and use of mobile phones change the way in which people and institutions interact, and modify accordingly the relation between personal and private space. The solutions to the privacy problems introduced or aggravated by the use of mobile phones cannot be rigid and final, but should rather be formulated as a personal attempt of every individual to dynamically balance his/her needs of privacy with the need to participate in networked social interactions.

REFERENCES

Altman, I. (1975). *The environment and social behavior: Privacy, personal space, territory and crowding.* Monterey: Brooks/Cole.

Altman, I. (1977). Privacy regulation: Culturally universal or culturally specific? *Journal of Social Issues, 33*(3), 66-84.

Broache, A. (2006, May 16). *Wireless location tracking draws privacy questions.* Retrieved May 2006 from http://news.zdnet.com/2100-1035_22-6072992.html

Evers, J. (2006, March 29). *Spy programme snoops on cell phones.* Retrieved April 2006 from http://news.com.com/Spy+program+snoops+on+cell+phones/2100-1029_3-6055760.html

Fortunati, L. (2002). The mobile phone: Towards new categories and social relations. *Information, Communication & Society, 5*(4), 513-528.

Geser, H. (2004). *Towards a sociological theory of the mobile phone.* Retrieved April 2006 from http://socio.ch/mobile/t_geser1.htm

Hulme, M., & Peters, S. (2001, April 1-2). Me, my phone and I: The role of mobile phones. *Proceedings of the CHI Workshop: Mobile Communications: Understanding Users, Adoption & Design,* Seattle, WA. Retrieved April 2006 from http://www.cs.colorado.edu/~palen/chi_workshop/papers/HulmePeters.pdf

Kindberg, T., Spasojevic, M., Fleck, R., & Sellen, A. (2004). *How and why people use camera phones.* Retrieved April 2006 from http://www.hpl.hp.com/techreports/2004/HPL-2004-216.pdf

Kristoffersen, S. (2005). *Privacy management for next generation mobile telephony.* Retrieved April 2006 from http://www.ifi.uio.no/forkning/grupper/is/wp/082005.pdf

Lettice, J. (2003, May 14). *UK gov seizes data on 100m calls, 1m users, a year.* Retrieved April 2006 from http://www.theregister.co.uk/2003/05/14/uk_gov_seizes_data/

McLeod, C. (2003). Sneaky cameras. *Press Council News, 15*(3). Retrieved April 2006 from http://www.presscouncil.org.au/pcsite/apcnews/aug03/cameras.html

Minch, R. P. (2004). Privacy issues in location-aware mobile devices. *Proceedings of the International Conference on System Sciences,* Hawaii. Retrieved April 2006 from http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/05/205650127b.pdf

Monk, A., Carroll, J., Parker, S., & Blythe, M. (2004). Why are mobile phones annoying? *Behaviour and Information Technology, 23*(1), 33-41.

Palen, L., & Dourish, P. (2003, April 5-10). Unpacking "privacy" for a networked world. *Proceedings of the CHI Workshop,* Ft. Laurendale. Retrieved April 2006 from http://ww.cs.colorado.edu/~palen/papers/palen-dourish.pdf

Perera, R. (2001, October 25). *Germany joins worldwide surveillance trend.* Retrieved April 2006 from http://www.itworld.com/Tech/2987/IDG011025surveill/

Plant, S. (2001). *On the mobile. The effects of mobile telephones on social and individual life.* Retrieved April 2006 from http://www.motorola.com/mot/doc/0/234_MotDoc.pdf

Richtel, M. (2005, December 10). *Live tracking of mobile phones prompts court fights on privacy.* Retrieved April 2006 from http://www.nytimes.com/2005/12/10/technology/10phone.html?ex=1291870800&en=2011ce3dd6b43183&ei=5088&partner=rssnyt&emc=rss

Townsend, A. M. (2000). Life in real-time city: Mobile telephones and urban metabolism. *Journal of Urban Technology, 7*(2), 85-104.

KEY TERMS

Biometric Security: A security science where body or physical attributes are used for secure identification and authentication, such as fingerprints, voice patterns, face geometry, hand geometry, retinal scans, signatures, and typing patterns.

Hacker: Individual who gains unauthorized access to computer systems for the purpose of stealing and corrupting data.

Intelligent Agent: A specialized software application that automatically collects information or performs another specific task on behalf of a person or a software program.

Location-Tracking Application: A virtual software application that permits third parties to identify the location of mobile phone users/owners.

Position-Aware Mobile Device: Information technology tools that are capable of defining their spatial location.

Spyware: Any technology that allows third parties to collect information about a person or organization without their knowledge and permission.

Ubiquitous Computing: A new information technology paradigm that attempts to integrate computation into the environment, rather than using computers that are distinct objects.

Mobile Phone Texting in Hong Kong

Adams Bodomo
University of Hong Kong, Hong Kong

INTRODUCTION: TECHNOLOGY AND LANGUAGE CHANGE

Mobile phone texting or communication through *short message service* (SMS) has emerged as a frequent daily linguistic, literacy, or general communicative practice in which two or more people exchange messages by coding and decoding texts received and sent from their cell phones. Mobile phone texting is almost now as pervasive and as ubiquitous as mobile phone voice communication. This communication process can be witnessed in buses, at homes, in offices, in restaurants, out in the woods, on the high seas, and even in the air! Hong Kong's main English language newspaper, the South China Morning Post (SCMP), on April 11, 2004, indicated that as huge a volume of 200 million SMS messages are exchanged monthly. SMS has become a multi-million dollar business for service providers.

Along with other kinds of digital technology-mediated communication, SMS seems to be causing a silent revolution with regards to the linguistic and communication habits of people in Hong Kong and beyond. This is especially so among the youth where one can safely say that more than 80% of people between the ages of 12 and 25 frequently use SMS as a mode of communication with their peers.

Given such a huge impact that this mode of communication has on the population, researchers and policymakers ought to turn their attention to this topic and attempt to find answers to questions about the consequences of SMS on various issues including language, communication, and our general social behaviors.

In this article, we focus on the relationship between communications technology and language change. Does the emergence of these new communications technologies affect our language and communication habits? Does it change our language, bringing in new words and structures of expressions, and does it alter our general communication patterns? In short, is technology changing our language?

In examining these questions, based on observation and analysis of issues of language, literacy, and communications technology, we propose a model called *technology-conditioned approach to language change and use* (TeLCU). This approach projects the view that there is a causal relationship between the emergence of new tools and media of communication and the creation of new forms of language and communication. New tools and media of communication demand the creation of new forms and ways of communication. These new forms compete with existing forms and ways of communication, leading to changes in the way we *use* language in its various forms, including spoken and written forms.

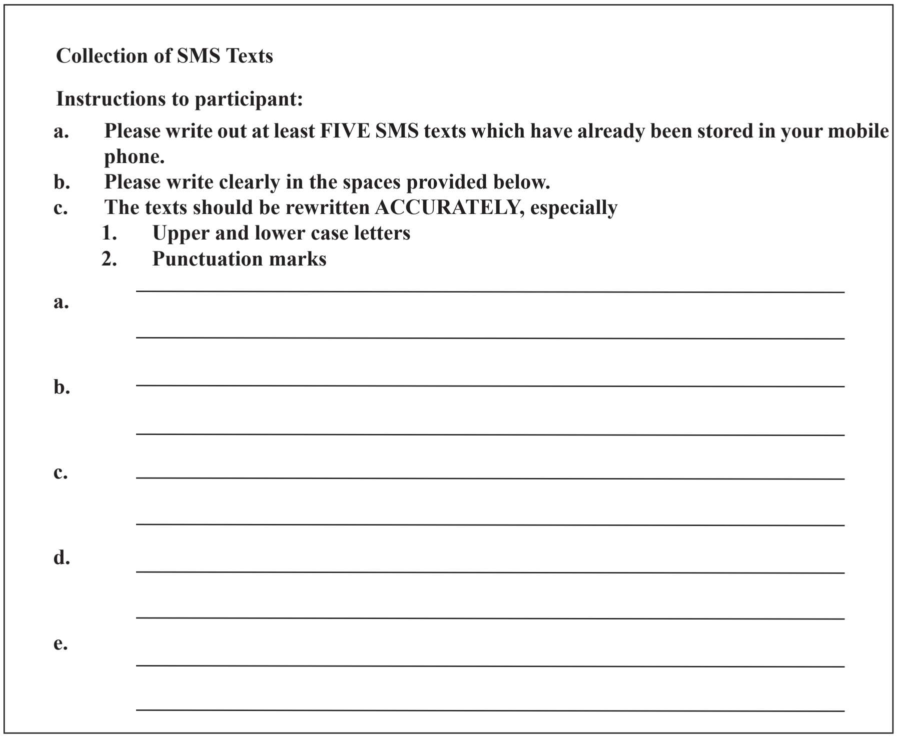
A potential anti-thesis to TeLCU is that there is little or no causal relationship between the prevalence of new media of communication and changes in the forms of language and ways in which we use human language. While not directly arguing against the idea that there is no causal relationship between the emergence of new technology, in general, and the new ways in which language is used, Kress (1998), for instance, observes that "… when we look at the far-reaching and deep changes in forms of communication which characterize the present-day e-mail and its changing forms of language, for instance, it is tempting to attribute these changes to some technological innovation but erroneous to do so" (p. 53). Luke (2000) also takes a similar position, believing that new forms of literacy practices do not simply emerge with technological change. Rather, "technologies always emerge as products of specific cultural practices, literate traditions, and the interests and desires of those groups who design and name them" (p. 83).

This article builds on this fruitful discussion in the literature on the relationship between new technologies and the way language is used within these technologies, and argues that there is indeed a significant causal relationship between communications technology and new language and communication practices or more specifically the evolution of new ways of using language. As Adams (1996) puts it, "the new technologies are themselves dramatically changing the nature of the language we use." Such an approach is also supported by Baron (1984), who concludes that "[n]o one in the computer industry has any hidden agenda for using hardware or software development to alter human language. Yet technology can indeed drive linguistic and social change" (p. 139).

Indeed, new practices of language and communication may be attributed to a set of unique properties in new communications technology. Modern digital communications technology is characterized by flexibility, connectivity, and interactivity (Blurton, 1999) that traditional forms of technology like radio and TV lack. In other words, it is possible to have many-to-many, many-to-one, one-to-many, and one-to-one modes of communication with modern digital

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Data collection form for data providers



Collection of SMS Texts

Instructions to participant:

a. Please write out at least FIVE SMS texts which have already been stored in your mobile phone.
b. Please write clearly in the spaces provided below.
c. The texts should be rewritten ACCURATELY, especially
  1. Upper and lower case letters
  2. Punctuation marks

a. ____
____

b. ____
____

c. ____
____

d. ____
____

e. ____
____

information and communications technology (ICT). These features of digital ICTs enable them to have a more pervasive influence on forms and uses of language.

In pursuing these questions and research agenda, this article is organized as follows. After giving background statistics and other kinds of information about the evolution of SMS, we introduce and describe our data which comprise a corpus of about 500 SMS messages collected from Hong Kong youth in April 2002. Following this, we introduce the theoretical model of TeLCU, sketching out the ways in which we believe new technologies affect our language structure and use. We then look closely at the linguistic features of the corpus, noting peculiar instances of structure and usage. The article concludes with an outline of the implications for these linguistic and communicative changes in our society with particular reference to social and educational habits, and consequences for the design of new tools of communication.

THE DATA: SMS CORPUS[1]

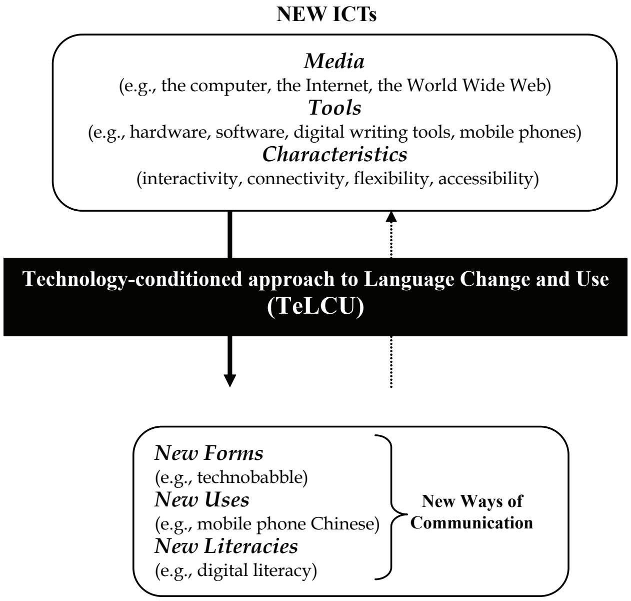
What is SMS?

Short message service, first introduced commercially in 1995, refers to the transmission of short text messages between mobile phone users. The first SMS message was a Christmas greeting sent out in Britain in 1992. Today, SMS has emerged as one of the major digital communication media, with an estimation of over one billion messages exchanged per day around the world. SMS had a slow beginning in Hong Kong, mainly because of the inconvenience of inputting Chinese characters with mobile phone keypads. In addition to that, text messages could only be exchanged between subscribers of the same service provider. However, the SMS market has grown rapidly in Hong Kong. In December 2001, the six major service providers opened their networks to allow message exchanges across networks (SCMP, 2002). As of April 2004, over 200 million messages are exchanged every month in the territory.

Data Collection

The collection of textual data began with a questionnaire survey conducted in April 2002. The target subjects of the survey were Cantonese-speaking youngsters in Hong Kong who use English as a second language. Since text messages involve private correspondences, prior permission was sought from the data providers. At the end of the questionnaire, the respondent was asked whether he/she would be willing to make his/her SMS texts available for the study. The study (Bodomo, 2001-2002) successfully collected 487 messages from 87 out of the 92 respondents who participated in the questionnaire survey. Data providers were asked to write

Figure 2. TeLCU—The relationship between new ICTs and new forms of language and literacy


NEW ICTs
Media
(e.g., the computer, the Internet, the World Wide Web)
Tools
(e.g., hardware, software, digital writing tools, mobile phones)
Characteristics
(interactivity, connectivity, flexibility, accessibility)
Technology-conditioned approach to Language Change and Use
(TeLCU)
New Forms
(e.g., technobabble)
New Uses
(e.g., mobile phone Chinese)
New Literacies
(e.g., digital literacy)
New Ways of Communication

down at least five messages that were already stored on their mobile phone message in/outbox on a form provided by the investigator (see Figure 1). Since the texts were collected from end users, the corpus may also include messages sent from a non-mobile device/tool, such as ICQ.

THE MODEL: TeLCU

In this section, as discussed earlier, we propose a model called *technology-conditioned approach to language change and use.* We suggest that there is a causal relationship between the emergence of new tools and media of communication in our society and the use of language.

New technologies often require new forms of language and literacy to express new concepts that emerge along with these new media and tools. New media of communication, then, can lead to changes in the way people use their language. In fact, works like Lakoff (1982) and Ong (1982) have discussed the impact of oral media like TV and radio on people's use of language.

The information age is characterized by the rapid advancement of technology, with the introduction of new tools and media. Among these, ICTs constitute the type of technology that this article is concerned with. ICTs, according to Blurton (1999), are characterized by flexibility, connectivity, and interactivity, which are different from passive, one-way media such as TV and radio. These distinctive features of ICT tools and media allow pervasive changes in language forms and uses. Apart from these, we believe that the popularity of a particular tool or media can also be a major factor in the discussion of how new ICTs introduce new forms and uses of language. One might expect that the products of the TeLCU model (i.e., new forms of language and literacy) are gradable. That is, the more of the above features a new technology carries, the more likely new forms of language and literacy will be introduced, and the more widespread these new forms will be. Figure 2 illustrates the cause-and-effect relationship between new digital ICTs and new forms of language and literacy.

In this model, we show that new ICTs include all the tools and media along with a bundle of features that act as the 'inputs' of this model. These features undergo the main process, TeLCU, and new forms of language and literacy are then generated. This is indicated with the downward arrow in the model. However, we do not eliminate the possibility of a 'reverse' process—that is, the possibility where new forms of language and literacy reinforce changes in new technologies. Hence, the upward arrow of the diagram illustrates this process. This process, nevertheless, cannot exist without TeLCU: new ICTs foster changes in language and literacy before new forms of language and literacy can have impact on ICTs.

The TeLCU model as espoused here is a specific model for a technologically deterministic approach to account for the linguistic and communication dynamics that are often observed in the environment of ICTs. As has been noted earlier, this technological determinism is not without its critics. Technological determinism is often juxtaposed with what is called socio-political determinism. Herring (2003) notes: "Many technologically deterministic claims regarding CMC, for example, have been found to make incorrect predictions in specific contexts of use, and a call has gone out for explanations of computer-mediated behaviour that take social and contextual factors into account" (p. 9). In response to this critical evaluation of technological determinism, we believe that there is hardly any need for this juxtaposition at all; indeed, the two approaches do not seem to be mutually exclusive, and certainly the TeLCU approach pursued here is not intended to exclude socio-political factors for language and communication dynamics in the age of IT. It may be true, for instance, that socio-political factors in Hong Kong such as a very high population density may force the youth of Hong Kong to constantly use mobile phone texting to sound each other out in busy malls and in large restaurants (as evidenced by the preponderant use of *nei5 hai2 bin1 aa3*, 'Where are you?' in both mobile phone voice and text messages); however, the crispness of the language they use would certainly be constrained by the technology available for decoding the messages.

Having introduced the theoretical framework of TeLCU, we shall discuss in the following sections some of the new forms of language and literacy under the influence of new ICTs.

LINGUISTIC FEATURES OF MOBILE PHONE TEXTING IN HONG KONG

As its name suggests, an SMS text message is indeed very short. It has to be short because users are constrained to input only up to 70 Chinese characters or 160 alphabetic letters. This also imposes some constraints on the basic structure of a message composed with the mobile phone. This section explores the various linguistic features identified in the corpus. The crisp nature of SMS language necessarily raises questions about its morphological and syntactic nature. Given a limited writing space and limited production time, there is logically more need to be as economical as possible than to be as elaborate/expressive as possible. In other words, a principle of economy outranks a principle of expressivity. The questions then are: How is this balanced out? What aspects of the morphology and syntax are retained, and what are given up? We here propose a general principle of economy and try to show how it generally manifests itself in SMS:

Technologically-Conditioned Economy of Expression

Words, phrases, and sentences should be coded with as few symbols as possible without giving up comprehensibility.

1. **Morphology:** Produce words as short as possible.
2. **Syntax:** Produce strings as short as possible.

Morpholexical Features: Shortening

We shall look at some morpholexical features in this subsection. While there are many morpholexical features of mobile phone texting that distinguish it from normal writing, we concentrate on the more salient feature of shortening. Shortening in ICT texts has been widely studied in the literature of computer-mediated communication (CMC). Scholarly research into SMS is still rare, and thus we have identified a set of shortened expressions in the corpus. We use the term here as a general cover word for the various strategies for shortening words, including acronymy, abbreviation, and so on. Table 1 is a list of common ways for shortening and abbreviating lexical items and the number of times this happens in the corpus.

The following example from the corpus illustrates the use of shortenings in the real messaging context:

r u 3 2nite? izit possible 2 hv dinner 2getda? call me (1)

Most (though not all) morpholexical items, as seen above, are thus produced in accordance with the economy principle: produce words as short as possible. We now turn to the syntactic aspects of the economy principle.

Syntax: Economy of Expressions vs. Expressivity

Although there exists a wealth of studies on the linguistic features of ICT texts, the aspect of *syntax* is rarely mentioned. This subsection is an attempt to take into account the syntactic nature of shortened expressions in SMS. The assumption here is that, in shortening a sentence, functional items such as tense and aspect are dropped, while lexical categories such as nouns and verbs are allowed to stay; in fact, this is more of a tendency than absolute adherence to the principle. So while in the morphology lexical items are more susceptible to reduction, in the syntax functional items are more susceptible to reduction or even dropping. Next we list each group of items likely to be retained or dropped.

- **What is retained**
    - lexical items
    - the most essential piece of information:
      lunch? (2)
      SLEPT? (3)
      QUIZ (4)
      exams finished la. now coming (5)
      in ikea,thinking of u... (6)
- **What is left out**
    - Pro-drop
      miss u terribly! wanna c u & hold u so much. will hv quite a bz wk. =((7)
      MISS U VERY VERY MUCH DUNNO HOW TO TELL U (8)
    - Tense
      YES I RECEIVE (9)
      My friend ask me to go taiwan (10)

The above are ways in which mobile texters attempt to satisfy the principle of economy over and above that of elaborate expressivity.

Phonological Features: Homophony

Some interesting morphophonological and phonosemantic features also characterize the language of mobile texting. In this article, I will focus mainly on the concept of homophony. One of the creative practices of mobile phone texters (Bodomo 2002) lies in their ability to substitute shorter (forms of) words for longer words that are homophonous or near-homophonous. While there are not that many examples in this particular corpus, we have found a very interesting one:

3 for ‘free’ as in

r u **3** 2nite? (11)

Here, the homophonous word *three* (its shorter form being the numeral 3) is used to represent *free*.

The Omission of Tone

Another of the crisp nature of mobile phone text is the explicit omission of tone markings in the Cantonese romanization. The Linguistic Society of Hong Kong (LSHK) has an elaborate romanization system, including tone markings[2], but this is hardly used in mobile texts. We list below a number of sentences showing various ways of romanization, especially in the rendition of final particles, which are very much a characteristic feature of the language of ICTs like mobile phone and ICQ texting (Lee, 2002a, 2002b).

exams finished la. now coming (12)

Where R U worry ar me (13)

Heavy raining wor... but u still need to go rite?
Be careful ar :p (14)

Romanization of Common Cantonese Expressions

Ai Ya... ho sun fu ar... the ref bk is very difficult...
can't understand a word ar! @-@ ho mun! (15)

Hey bei sum gei la... (16)

Lo lic... everything will be fine again (17)

Emoticonymy

Unlike face-to-face talk where emotions can be expressed visually (by facial expressions) or auditorily (by stress and loudness), text-based CMC mainly takes place without any face-to-face interactions between participants. Therefore, emotions are often expressed by graphical means. The introduction of the so-called *smileys* or *emoticons* has facilitated electronic communication to a large extent.

By emoticonymy*,* we mean "a subfield of CMC which involves the analysis and practices of employing smileys and related icons for conveying emotions and other linguistic and kinesic features intended by the author" (Bodomo & Lee, 2002, p. 35). Emoticons (emotional-icons) are also used to avoid misunderstanding in the interpretation of a message. For instance, a smiling face like :) is often used to indicate a joke or that the message sender is happy about something. Emoticons are usually placed at the end of a sentence, as in "I am so sorry :-(." Most emoticons are created with different combinations of punctuation marks (colon and a closed bracket form a smiling face). This is further illustrated as follows:

Ladies hockey champion! Yeah >:D (18)

Tmr 記得 交essay ^v^ (19)

tmr gei3 dak1 gaau1 essay (20)

tomorrow remember submit essay (smiles) (21)

Don't forget to hand in your essay tomorrow. (22)

Mode-Mixing

Apart from emoticons, there also exist combinations of codes and symbolic systems in the messages. As Kress (2000) suggests, 'The appearance of modes other than language in the centre of the domain of public communication has several aspects: new, or newly prominent modes appear: texts, textual objects are more clearly seen to be multimodal, that is, to be constituted by a number of modes of representation" (pp. 183-184). So what is happening in SMS? In our corpus, we can identify a combination of words/letters and numbers, and graphic and alphabetic symbols. These are illustrated as follows:

- Words/letters + Numbers

r u 3 2nite? izit possible 2 hv dinner 2getda?call me
(Are you free tonight? Is it possible to have dinner together? Call me.) (20)

- Graphical + alphabetic

( )""( )

(▌'▌o▌'▌)")

(.) /

Happy New Year

(Punctuation marks to visualize 'fireworks') (21)

U :)= I :) U :(= I :(+ WORRY TAKE CARE

(When you're happy, I am happy too. When you are sad, I also feel sad and worried. Take care.) (22)

New Varieties of English

In this section we have examined specific features of written language on the mobile phone. The constraints of time and space force users to invent diverse and new ways of shortening or even leaving out entire words, and adding new signs and icons to make themselves comprehensible.

We believe that the media of the mobile phone and other ICTs, such as the ICQ that is popular in Hong Kong, are changing our language use in profound ways, and making Hong Kongers and the youth in other parts of world evolve different ways of using language. New varieties of written English thus seem to be evolving. This new variety involves a mixture of English and Chinese signs, characters, words, and expressions. Even though some features may seem spontaneous and playful at first glance, some features and patterns are used quite consistently and repeatedly by different users. Indeed, these new varieties of written English are finding their way into school writing and are the source of worry for school teachers, but this is the topic for another study (Bodomo, 2002).

SUMMARY AND CONCLUSION

In this article, we have attempted to take stock of language structure and use within the technological environment of new media of communication such as the mobile phone. Drawing from different arguments and points of view within the literature, we have highlighted the important debate as to whether the linguistic and communicative transformation that we are witnessing is one of *technological determinism* or *social determinism.* The position espoused in this article is that while technological determinism cannot be divorced from social determinism, one has to tease out the two and look concretely and specifically at the immediate ways in which language functions and changes in the environment of new ICTs. We have proposed a specific technologically deterministic model known as TeLCU. As has been proposed, users of language are forced to abide by a quite stringent *economy principle* in mobile phone texting. This has resulted in different ways of experimenting with language structure and use—phonologically, morphologically, and syntactically. The attendant innovations at these various levels of linguistic analysis seem to be creating new varieties of language and new ways of communicating.

Modern ICTs such as computers, the Internet, and mobile phones seem to be profoundly transforming our language structure and general communicative patterns so much so that linguists and other language and communication professionals who turn a blind eye to this (r)evolution risk being accused of professional negligence.

REFERENCES

Adams, A. (1996). Language awareness and information technology. *Curriculum and Teaching, 11*(2), 69-76.

Baron, N.S. (1984). Computer mediated communication as a force in language change. *Visible Language, XVIII*(2), 118-141.

Blurton, C.G. (1999). *New directions of ICT-use in education.* UNESCO's World Communication and Information Report 1999. Retrieved from http://www.unesco.org/education/educprog/lwf/dl/edict.pdf

Bodomo, A. (2000). *The Hong Kong ICT corpora.* University of Hong Kong.

Bodomo, A. (2001-2002). *Linguistic features of mobile phone communication.* CRCG Research Project, HKU. Retrieved from http://www.hku.hk/linguist/research/bodomo/MPC

Bodomo, A. (2002, May 11-12). Communicating in the age of information technology: New forms of language and literacy and their educational implications. *Proceedings of the Interactive Seminar & Discussion,* Ansted University, Malaysia.

Bodomo, A. (2003-2005). *Communicating in the age of information technology: New forms of language and their educational implications.* Research Project Funded by Quality Education Fund (Ref: 2002/0384). Retrieved from http://www.hku.hk/linguist/research/bodomo/QEF/

Bodomo, A., & Lee, C. K. M. (2002). Changing forms of language and literacy: Technobabble and mobile phone communication. *Literacy and Numeracy Studies, 12*(1), 23-44.

Herring, S. C. (2003). Media and language change: Introduction. *Journal of Historical Pragmatics, 4*(1), 1-17.

Kress, G. (1998). Visual and verbal mode of representation in electronically mediated communication: The potentials of new forms of text. In I. Snyder (Ed.), *Page to screen: Taking literacy into the electronic era, 1998* (pp. 53-79). London/New York: Routledge.

Kress, G. (2000). Multimodality. In B. Cope & M. Kalantzis (Eds.), *Multiliteracies: Literacy learning and the design of social futures* (pp. 182-202). London; New York: Routledge.

Lakoff, R. T. (1982). Some of my favorite writers are literate: The mingling of oral and literate strategies in written communication. In D. Tannen (Ed.), *Spoken and written language* (pp. 239-260). Norwood, NJ: Ablex.

Lee, C. K. M. (2002a). *Chinese and English computer-mediated communication in the context of new literacy studies.* MPhil thesis, The University of Hong Kong.

Lee, C.K.M. (2002b). Literacy practices in computer-mediated communication in Hong Kong. *The Reading Matrix*, *Special Issue: Literacy and the Web.* Retrieved from http://www.readingmatrix.com/articles/lee/article.pdf

Luke, C. (2000). Cyber-schooling and technological change: Multiliteracies for new times. In B. Cope & M. Kalantzis (Eds.), *Multiliteracies: Literacy learning and the design of social futures* (pp. 69-91). Australia: Macmillan.

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word.* London/New York: Routledge.

South China Morning Post. (2002). *Council urges push on SMS.* October 17.

South China Morning Post. (2004). *Text maniacs find their calling.* April 11.

KEY TERMS

Digital Literacy: "The ability to understand and use information in multiple formats from a wide range of sources when it is presented via computers" (Gilster, 1997, p. 1).

Hong Kong ICT Corpora: Developed at the University of Hong Kong, the Hong Kong ICT corpora is a database of texts collected from various modes of ICTs, including computer-mediated communication such as ICQ and e-mail, and mobile phone texting.

Information and Communications Technology (ICT): The use of technology to communicate and process information.

Literacy: The ability to code and decode linguistic and other symbolic systems for communication and information processing (Bodomo, 2000).

Mobile Phone Texting: The transmission of short text messages between mobile phone users.

Short Message Service (SMS): See *Mobile Phone Texting.*

Technobabble: A language characterized by the pervasive use of technical jargon, or a speech or a piece of writing characterized by the pervasive and extreme injection of technical jargon to the extent that the language is barely comprehensible to non-specialist speakers of the particular language.

Technology-Conditioned Approach to Language Change and Use (TeLCU): The central claim of TeLCU is that there is a causal relationship between the introduction of new tools and media of communication and the emergence of new forms of language and communication.

ENDNOTES

[1] This SMS corpus is part of the Hong Kong ICT-corpora (Bodomo, 2000) developed at the University of Hong Kong. Some texts of the corpora are available online from http://www.hku.hk/linguist/research/bodomo/QEF/

[2] Tones in Cantonese are marked as follows: Tone 1: high level, Tone 2: high rising, Tone 3: mid level, Tone 4: low falling, Tone 5: low rising, and Tone 6: low level.

Mobile Phones for People with Disabilities

Hend S. Al-Khalifa
Southampton University, UK

AbdulMalik S. Al-Salman
King Saud University, Saudi Arabia

INTRODUCTION AND BACKGROUND

Nowadays, the demand for mobile phones by people with special needs is evolving. Disabled people can utilize mobile phones for personal communication, security, social integration, and autonomy. Personal communication is one of the most important reasons that disabled person uses a mobile phone. For example, people with motor disabilities cannot easily reach the wired telephone within a limited period of time when it rings. Security is another reason that most disabled people strive for. In case of emergency, illness, or accident, mobile phones are considered a fast communication channel. Furthermore, to keep in touch with society and to feel autonomous, people with special needs think that the mobile phone is a good medium for social integration and self-independence (Abascal & Civit, 2000).

Enrico and Stephen (2003) stated that mobiles could also be used as an aid to carry out everyday activities, as though they were not disabled—for example, using the mobile phone to remotely instruct PCs, lifts, doors, and ATMs. Abascal and Civit (2000) emphasize Enrico and Stephen's (2003) argument on the use of mobile phones to control other devices by saying "with the emerging Bluetooth standard, a user will be able to use her/his mobile phone controlling it from any other device including not only PDAs and notebooks but also Assistive Technology devices such as communicators and wheelchair controllers" (Abascal & Civit, 2000, p. 264).

Common problems of existing mobile phones arise from small print on mobile phone controls and screens, hard-to-press buttons, complexity of use, no audio battery limit indicator, no caller identification, and no specially designed keys for emergency and easy access to specific functions (Lee, 2003). Another problem is the limitation of some software programs that only work on specific phone brands and operating systems. Also, there is no standardization between mobile phone companies in terms of design and functionality. Furthermore, as stated by Baker and Bellordre (2004), the barriers to access/use of mobile phones can be classified into three factors: awareness and proficiency factors, economic barriers, and incompatible technologies. Therefore, the lack of standardization imposes the proposal of strict rules and regulations for designing accessible telecommunication devices by the Federal Communications Commission (FCC); these rules are called "Telecommunications Act of 1996, Section 255" (Chen, 1999; Telecommunications Act of 1996[1]). The FCC has rules requiring telecommunications manufacturers and service providers to make their products and services accessible to people with disabilities, if readily achievable. These rules implement Section 255 of the Communications Act. Where it is not readily achievable to provide access, Section 255 requires manufacturers and providers to make their devices and services compatible with peripheral devices and specialized customer premises equipment that are commonly used by people with disabilities, if such compatibility is readily achievable.

Finally, many disability-centered organizations worldwide perform regular reviews and tests on the latest mobile phones in the market and publish the result online for free. As an example, *Accessworld,* a publication of AFB (American Foundation for the Blind), is carrying out regular mobile phone reviews and evaluations to test the accessibility and usability of several off-the-shelf cellular telephones and add-on software applications. These reviews can be accessed online at http://www.afb.org/.

TECHNOLOGIES, SERVICES, RESEARCH, AND PROJECTS

Using mobile phones is challenging for the movement impaired and physically challenged, the blind and visually impaired, the elderly and arthritic, and those with any of many immune and neuromuscular disorders. In order to make mobile phone technology accessible for those people, this may require augmentative communication devices with expensive customized hardware and software interfaces to support their interaction with mobile phones. Moreover, there is a need for a standard universal design of cell phones that includes features such as infrared ports, volume range, speakerphones, matrix displays, EZ buttons, voice dialing, and messaging (Bryen, 2004). Also, there are many services and much research currently taking place to bridge the gap of the mobile divide. In this section, we will present the various trends and technologies that are developed to help people with special needs cope with the new stream.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Owasys 22C mobile phone

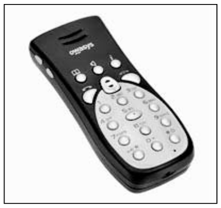
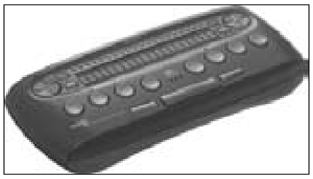


Figure 2. ALVA MPO



Technologies

Technologies used in mobile phones can be classified into three types: software technologies, which will handle all aspects of integrating software functionality into a mobile phone; hardware technologies, which will focus on the different devices that can be added to a mobile phone to make it more accessible and the variant kinds of newly designed mobile phones devoted for the use by people with special needs; and hybrid technology, which combines both software and hardware into a single device.

Software Technologies

Most mobile phones that operate using the Symbian operating system can easily install third-party software like screen readers and screen magnifiers, which provide further accessibility to mobile phones. On the other hand, off-the-shelf mobile phones have limited speech output capabilities (Burton, 2005). In this subsection, we provide the reader with accessible software applications that work on Symbian-based mobile phones.

SpeechPAK TALKS is a product from ScanSoft[2] Company. It converts the display text of a cellular handset into speech, making the phone accessible for visually impaired people. SpeechPAK TALKS runs on Symbian-powered mobile phones. It speaks to the user either in English, German, or another language. The user can change ring tones for different callers, check who dialed the number, hear spoken voice messages, write and send an e-mail or a fax, and manage PDA functions. A portable Braille display can also be attached.

Mobile Speak[3] is another screen reading software that can be installed on a mobile phone. It is available in many languages including Arabic.

Talks[4] is also a screen reader for mobile phones. It supports Arabic and English languages. A trial version of this software is available through the Internet.

Mobile Magnifier[5] is language-independent software that enlarges and enhances all items of the mobile phone display. It provides six different color schemes (from black and white up to 4096 colors). Mobile Magnifier automatically detects and magnifies the area of interest as the user navigates through the phone's user interface. It supports a range of mobile phone brands like Nokia and Siemens.

The vOICe MIDlet for Mobile Camera Phones[6] is seeing-with-sound technology for the totally blind. It is available for most camera phones and PDAs. The vOICe MIDlet software runs on both Symbian and non-Symbian devices, and it is free of charge and can be downloaded from the Internet.

Mobile Color Recognizer (MCR)[7] is software that was developed to work with Mobile Speak and Mobile Accessibility. It can be installed in camera phones that are Symbian complaint. MCR can be used to determine the color (or different colors) of an object by taking its picture. MCR also can be used to know the level of light.

Hardware Technologies

Many major telecommunication players in the market have a department devoted to accessibility research. Nokia, Sony Ericsson, and Samsung, to name just a few, are all working on making their products more accessible for various types of disabilities. In this subsection, we provide the reader with accessible hardware technologies that can be used as a replacement/companion to a mobile phone and are tailored to the needs of a person with a disability (i.e., blindness, deafness, or motion impairments).

Owasys 22C phone[8] (Figure 1) is a cell phone that is designed specifically for blind and partially sighted people. The phone does not include a visual display; instead it uses speech synthesis technology to read everything that would normally appear on the screen. It has an ergonomic design with tactile keys. It also has a dedicated key to access the phonebook and an information key that verbalizes the user position within the phone's menu system. It gives audio feedback from the press of a button.

ALVA MPO[9] (Figure 2) is a Braille-based mobile phone. Besides making calls using the phone, it also has many integrated functionalities like a text-to-speech engine, organizer, and note-taker.

Mobile Screen Magnifier[10] (Figure 3) is a piece of a magnifier lens that is placed on top of a mobile display to magnify the text displayed on the screen up to 100%.

Voice activation/recognition is built into most mobile phones. A blind person can activate the mobile phone by dictating commands. Yet, a major disadvantage of this

Figure 3. Mobile screen magnifier

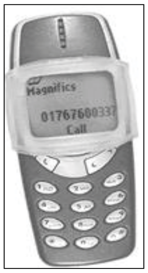


Figure 5. Mobile phone attached to TTY device

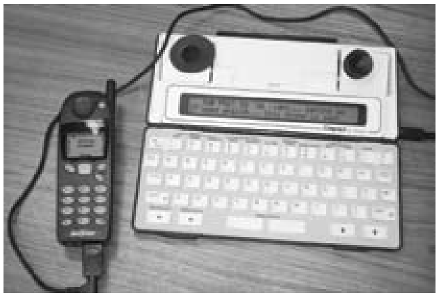
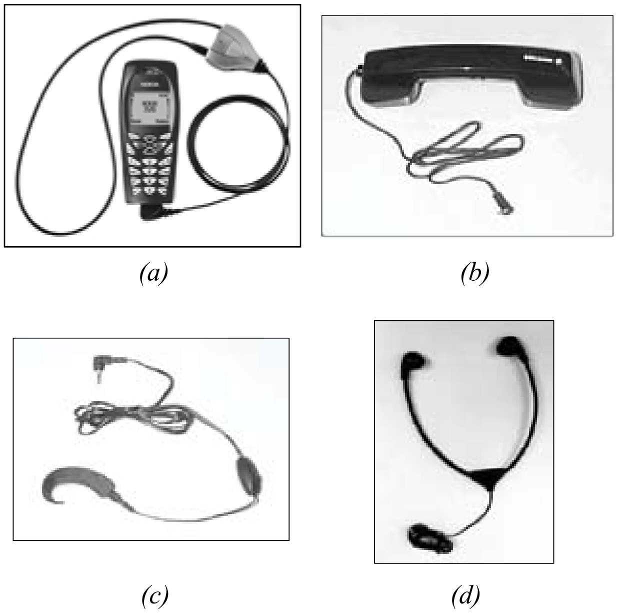


Figure 4. From (a)-(d): Nokia Loopset, SonyEricsson Handset Adapter, HATIS, and HITEC



(a) *(b)*

(c) *(d)*

functionality is the interference of the background noise and the battery power requirements. The Samsung P207 phone is a new phone that uses voice-recognition technology to allow users to create text messages by simply speaking into the phone.

Most telecommunication companies like Nokia and SonyEricsson have designed some of their mobile phones with a feature of amplifying or attenuating the sound frequency and volume to fit the hearing-impairment needs—that is, *adjustable audio.* Such phones, like the Nokia 3300, have this built-in feature.

Almost all mobile phones in the market have some sort of vibrating sense and flashing displays feature for helping hearing-impaired people to have some kind of *tactile and visual feedback.*

Hearing accessories[11] (Figure 4) include, for example, Loopset, which is a special device that someone can wear around his or her neck with built-in microphone for hands-free operation. It prevents the mobile phone from interference with the hearing aid used by hard-of-hearing people. Another accessory from SonyEricsson is the cellular phone handset adapter that links mobile phones with 2.5mm jacks to any TTY/TDD that has acoustic cups. HATIS (Hearing Aid Telecommunications Interconnect System) from SonyEricsson is another accessory that includes a silhouette induction loop that connects with a T-Coil in a hearing aid. It also includes a throat microphone that helps hands-free operation. In addition, the HITEC under-the-chin binaural headset works in a similar way as HATIS, with binaural sound for improved hearing and a microphone incorporated into the "Y" junction.[12]

TTY[13] (Figure 5) an acronym for teletype (also known as Telecommunications Device for the Deaf or TDD device). It is a special teletypewriter device with a screen to display typed text electronically that is connected to the mobile phone to type and read the conversation during a call, rather than using speech and hearing. Furthermore to benefit from these devices there must be telecommunication relay services (TRS is an operator service used by deaf persons and those with other hearing difficulties, to allow them to place telephone calls as voice messages by a TRS operator, and vice-versa) that support this technology.

Another available hardware technology is the *snap-on/wireless keyboard*[14] (Figure 6). Nokia wireless QWERTY keyboard is a small, foldable, lightweight keyboard that operates using A3 batteries. It is designed for disabled people who find pressing the buttons on the mobile phone awkward. A similar product is the Sony Ericsson T28 snap-on keyboard (Figure 6).

Hands-free accessories (Figure 7) are usually used by car drivers to make driving safer. On the other hand, people with motion impairments can benefit from such a technology.

The *voice-activated communication system*[15] Liberty Bell VAS allows individuals with minimal or no use of their hands the ability to communicate without any assistance. The user may initiate a call and answer a call, as well as adjust the volume simply by speaking into the mobile phone.

Figure 6. Nokia Wireless keyboard (left) and Ericsson T28 with an attachable QWERTY keypad (right)

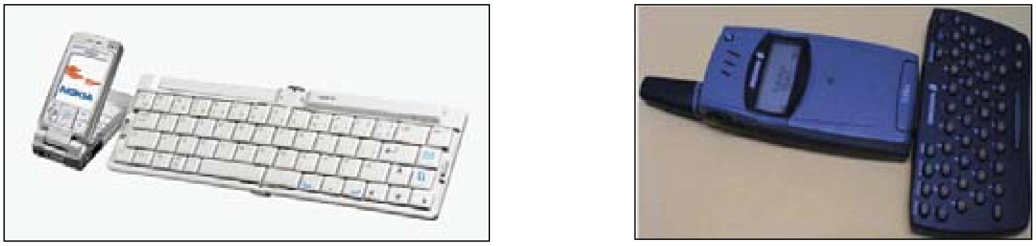


Figure 7. Bluetooth ear set

Figure 8. RNID mobile textphone

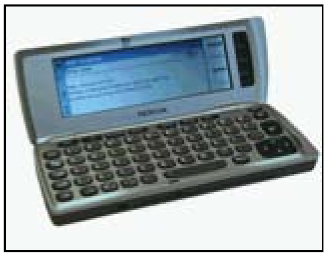


Figure 9. BlackBerry mobile phone

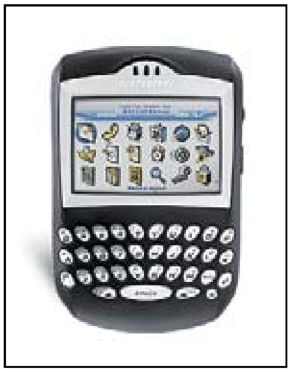
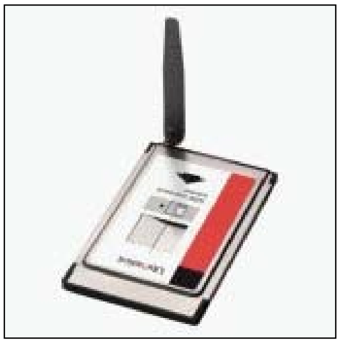


Figure 10. Ubinetics GC201 GSM PCMCIA phone card for laptops



Hybrid Technologies

Hybrid technologies are those technologies that integrate both hardware and software on the same device and designed for the people with special needs. This subsection lists some of these technologies.

RNID Mobile TextPhone[16] (Figure 8) is a combination of the Nokia Communicator 9210i and RNID's (the Royal National Institute for the Deaf's) mobile textphone software. The device enables deaf, hard of hearing, or speech-impaired people to make or receive telephone calls in text format.

The *Blackberry 7290 Wireless Handheld*[17] (Figure 9) is a mobile phone with many advanced features like e-mail, SMS, Bluetooth support, and organizer applications. Also, it has an integrated QWERTY keyboard. This device is capable of communicating with TTY or TDD terminals.

The *Ubinetics GC201 GSM PCMCIA phone card for laptops*[18] (Figure 10) can help a disabled person to deal with laptops instead of mobiles. Therefore, he/she can make calls; send and receive e-mails, faxes, and SMS messages; as well as browse the Internet. This card is a dual-band GSM Type II PC card, supporting voice, data, fax, and SMS. It can be used within any laptop that has a PC card slot and supports MS-Windows.

SERVICES

Many telecommunication companies support a wide range of services to help people with special needs cope with the digital divide. TTY, GPS, SMS, MMS, WAP, and mobile e-mail are types of services that can be used by both regular people and those with special needs (Høeg, Rasch, Ruud, Jynge, & Zäll, 2004).

TTY service is devoted mainly to deaf people. To benefit from this service, four groups are responsible for making TTY service worthwhile: the wireless network provider in a country, the mobile phone manufacturers, wireless service providers who provide the connection between the mobile phone and network system, and the TTY manufacturers who modify their TTYs so they can work with the handsets (Harkins & Barbin, 2002).

Mobile phones that are equipped with built-in or external GPS (global positioning system) receivers can pick up navigation signals from satellites and convert these signals into an accurate position in the phone (Makino, 1996). By calling the phone, its position is returned to a control center, where the exact position of the phone can be detected. The Nokia 9210/9210i with GPS module is equipped with a graphic color screen that shows the user's position. This type of phone is ideal for wheelchair users and others who need to call for help and tell people exactly where they are. Moreover, these phones can be used as a personal safety alarm (Høeg et al., 2004).

SMS (short messaging service) is a service for sending and receiving text messages on all modern mobile phones. A new trend is to use a phone with predictive text. This service is called T9.[19] This means that the phone 'guesses' which word the user is trying to type, which can make it easier and faster to use. This service is useful for deaf and motion-impaired people.

MMS (multimedia messaging service) is similar to SMS; it allows sending text, speech, images, sound, and video clips from one mobile to another. MMS also supports e-mail addressing, so that messages can be sent by e-mail. A mobile phone with WAP facility enables accessing information and services stored in WAP format on the Internet (Høeg et al., 2004; Ovum[20]).

WAP (wireless access protocol) is designed to access the Internet from mobile phones and PDAs using wireless communication. The use of the Internet is expected to increase rapidly even among disabled and elder people (Abascal & Civit, 2000).

MobileMail, a type of *mobile e-mail,* can be accessed via the Internet, WAP, SMS, or a mobile answering service. MobileMail can be printed out as a fax, or read out from the mobile answering service (mobile answering service e-mail). When activating the mobile answering service e-mail, the entire e-mail can be read out, including the sender name and the subject (Høeg et al., 2004).

RESEARCH AND PROJECTS

Since the introduction of mobile phones, many research and projects were conducted to potentially benefit from the different features a mobile phone can provide (Fernandez & Roa, 2003). As an example, in Japan, Katuhiro and Iwao (2004) from Tokyo Gakugei University, used a mobile videophone to deliver e-learning service for a student with severe physical impairment. Also, Fujitsu's company designed an easy-to-use mobile phone with a universal design and user-friendly interface that makes full use of speech synthesis and voice recognition technologies. The phone is called Raku Raku PHONE and is dedicated for elderly persons, persons unfamiliar with mobile phone operation, and persons with physical disabilities, for example, persons with visual disabilities (Toru, Keigo, & Yukinori, 2005).

In Slovenia, a project called GOVOREC is a system to automatically convert any Slovenian text into speech. "Currently, several leading Slovenian telecommunication companies are testing the system for providing information (e-mail, short messaging service—SMS, weather reports, traffic information) through mobile phones" (Sef & Gams, 2003, p. 227).

In the United States, "Mobile Wireless Technologies for Persons with Disabilities," a dedicated center at Georgia Institute of Technology, has been established to design wireless aids that target different disabilities, including mobility, vision, and hearing. It also tries to influence wireless-device manufacturers to make their existing products accessible to disabled people (Anonymous, 2004).

In Finland, there is ongoing research by VVT Institute about using video messaging as a way of communication. Kasesniemi, Ahonen, Kymaelaeinen, and Virtanen (2003) state that video messaging was the first possibility for the deaf to use signing in mobile communications, in their own language.

In Spain, the CONFIDENT project, partially funded by the European Commission, is a system to support disabled people managing their daily life by means of IT-based services. The platform of the system is based on the so-called "access from anywhere paradigm" and is built on a distributed and collaborative architecture based on Web services. This structure provides secure access to a service network from various devices for people with special needs (Cabrera, Arredondo, & Villalar, 2004).

In Europe, a consortium of researchers from Germany, Spain, Sweden, and the UK are working together on a project called WISDOM (Wireless Information Services for Deaf People on the Move). This project will enable hearing-impaired users to call up news, weather, and sports information in sign language from a video server via 3G (third-generation) phones, give commands to their phones in sign language, and access a real-time interpretation service to aid them in communicating with hearing people (Britta & Karl-Friedrich, 2001)

Much research has been conducted to discuss mobile phone design issues for people with special needs. As an example, Smith-Jackson, Nussbaum, and Mooney (2003) developed and used a framework called NARA (needs analysis and requirements acquisition) based on the universal design principles, to help suppliers and telecommunication companies construct mobile phones tailored to the requirements of people with disabilities.

The largest obstacle to mobile phone use by people with severe disabilities is the lack of a suitable interface system that would allow the user to operate the phone. Building interfaces to make mobile phones more accessible is another important theme that the augmentative and alternative communication (AAC) research project is tackling. The research is trying to build an interface system that bridges the gap for AAC device users and their mobile phones. A prototype interfacing system was demonstrated at the Australian Rehabilitation & Assistive Technology Association National (ARATA) Conference 2004 (Nguyen, Garrett, & Downing, 2004)

A personal navigational aid attempts to address many of the problems faced by persons who are blind. This aid incorporates many of the features of a PDA, cellular phone, wireless Internet access, and GPS (Szeto, 2003).

Finally, many guides have been compiled to help disabled people choose the suitable mobile phone with accessibility features; examples include those by Høeg et al. (2004) and Nguyen (2002).

CONCLUSION

As new technologies emerge and as the number of users grows, it is incumbent on technology developers to enhance their systems to help people with disabilities. One of the most important technologies in this respect is the mobile phone. There are many international worthwhile efforts to enhance the use of technology to people with special needs. This survey shows the great efforts from technology developers to assist people with different types of disabilities. In this article, we presented an overview of the existing hardware/software technologies that improve disabled people accessing mobile phones. The review takes into account the various developments oriented to people with disabilities, including software technologies, hardware technologies, hybrid technologies, services, research, and projects.

However, still much work can be done to improve the functionality/operation of mobile phones for people with special needs. This might include making mobile phones smarter, by tracing eye movements as an indicator of a dialing pattern, enabling the use of breath (puff and sip) controllers for people with movement impairments, operating a mobile phone by brain (e.g., dialing memorized numbers), or even enhancing the overall design of the handset to fit the disabled needs. So, as mobile phones are getting smarter and smarter everyday, we might soon see people with disabilities fully immersed in the telecommunication world.

REFERENCES

Abascal, J., & Civit, A. (2000). Mobile communication for people with disabilities and older people: New opportunities for autonomous life. In P. L. Emiliany & C. Stephanidis (Eds.), *Information Society for All, Proceedings of the 6th ERCIM Workshop on User Interfaces for All* (pp. 255-268). Consiglio Nazionale delle Ricerche, Firenze.

Anonymous. (1999). Phones for the disabled. *Rural Telecommunications, 18*(2), 9.

Anonymous. (2003/2004). Wireless for the disabled. *Technology Review, 106*(10), 64-67.

Baker, P. M. A., & Bellordre, C. (2004). Adoption of information and communication technologies: Key policy issues, barriers and opportunities for people with disabilities. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04).*

Britta, B., & Karl-Friedrich, K. (2001). Towards a 3rd generation mobile telecommunication for deaf people. *Proceedings of the 10th Aachen Symposium on Signal Theory* (pp. 101-106).

Bryen, D. N. (2004). Augmentative and alternative communication and cell phone use: One off-the-shelf solution and some policy considerations. *Assistive Technology, 16*(1), 11-17.

Burton, D. (2005, March). You get to choose: An overview of accessible cell phones. *Accessworld, 6*(2). Retrieved April 20, 2005, from http://www.afb.org/afbpress/pub.asp?DocID=aw060206

Cabrera, M. F., Arredondo, M. T., & Villalar, J. L. (2004, May). Mobile systems as a mean to achieve e-inclusion. *Proceedings of IEEE Melecon* (pp. 653- 656).

Chen, K. (1999). FCC to issue rules to help the disabled in the use of telecommunications gear. *Wall Street Journal (Eastern ed.),* (July), 1.

Enrico, B., & Stephen, K. (2003). Mobile devices: Opportunities for users with special needs. *Lecture Notes in Computer Science, 2795,* 486-491.

Fernandez, F., & Roa, L. (2003). Adaptive telecommunication system for disabled people. *Annals of Telecommunications, 58*(5-6), 890-904.

Harkins, J., & Barbin, C. (2002, May). *Cell phones get smart—And more accessible.* Retrieved April 20, 2005, from http://tap.gallaudet.edu/WirelessTelecom/AccessibleCell.htm

Høeg, M., Rasch, B., Ruud, S. E., Jynge, V., & Zäll, C. (2004). Mobile telephony for people with disabilities—A guide to choosing a mobile phone. *The Nordic Forum for Telecommunication and Disability*.

Kasesniemi, E.-L., Ahonen, A., Kymaelaeinen, T., & Virtanen, T. (2003). Elavan mobiilikuvan ensi tellenteet: Kayttajien kokemuksia videoviestinnasta (Moving pictures: User experiences about video messaging). *VTT Tiedotteita—Valtion Teknillinen Tutkimuskeskus, 2204,* 3-95. (In Finnish).

Katuhiro, K., & Iwao, K. (2004). Mobile videophone and e-learning for students with physical impairments. *Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'04)* (p. 203).

Lee, A. (2003). Breakdown communications. *Engineer, 292*(7621), 19.

Makino, H. (1996). Development of navigation system for the blind using GPS and mobile phone combination. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology* (Vol. 2, pp. 506-507).

Nguyen, T. (2002). Accessible mobile phone project. *ARATA News, 28,* 1-2.

Nguyen, T., Garrett, R., & Downing, A. (2004). Mobile phone access via an augmentative and alternative communication device. *Australian Rehabilitation & Assistive Technology Association National.*

Sef, T., & Gams, M. (2003). SPEAKER (GOVOREC): A complete Slovenian text-to-speech system. *International Journal of Speech Technology, 6*(3), 277-287.

Smith-Jackson, T., Nussbaum, M., & Mooney, A. (2003). Accessible cell phone design: Development and application of a needs analysis framework. *Disabil Rehabil, 25*(10), 549-560.

Szeto, A. J. (2003). A navigational aid for persons with severe visual impairments: A project in progress. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology* (Vol. 2, pp. 1637-1639).

Toru, I., Keigo, M., & Yukinori, N. (2005). Universal design activities for mobile phone: Raku Raku PHONE. *Fujistu Scientific and Technical Journal, 41*(1). Retrieved July 20, 2005, from http://www.fujitsu.com/global/news/publications/periodicals/fstj/archives/vol41-1.html

KEY TERMS

Braille: A writing system using a series of raised dots to be read with the fingers by people who are blind or whose eyesight is not sufficient for reading printed material.

EZ Button: A special button on a mobile phone that allows the reading of the label for any key, as well as the contents of the display and all menus and features of the phone.

MIDlet: A Java program for embedded devices, more specifically the J2ME virtual machine. Generally, these are games and applications that run on a cell phone.

Personal Digital Assistant (PDA): A small handheld computer that acts as a personal organizer.

QWERTY Keyboard: A name for the standard computer keyboard, named by the first six keys from the left on the top alphabetic row.

TDD Device: Stands for Telecommunications Device for the Deaf; allows a person to transmit typed messages over phone lines to another person with a TDD.

TTY: Abbreviation for teletypewriter. Machinery or equipment that employs interactive text-based communications through the transmission of coded signals across the telephone network.

ENDNOTES

[1] http://www.fcc.gov/telecom.html

[2] http://www.scansoft.com/speechpak/

[3] http://www.quantech.com.au/products/other_products/Screen_Reading/mobile_speak.htm

[4] http://www.nattiq.com/index.asp

[5] http://www.codefactory.es/

[6] http://www.seeingwithsound.com/midlet.htm

[7] http://www.independtech.com/accessories/morecognize.html

[8] http://www.owasys.com

[9] http://www.alvampo.com/

[10] http://www.mabels.org.uk/helpfulgadgets.htm

[11] http://www.nokiaaccessibility.com/

[12] http://www.ericsson-snc.com

[13] http://en.wikipedia.org/wiki/Telecommunications_Relay_Service

[14] http://www.nokiaaccessibility.com/

[15] http://www.planetmobility.com/store/phones/

[16] http://www.vodafone.co.uk/specialneeds

[17] http://www.blackberry.com

[18] http://www.expansys.com/product.asp?code=UBINGC201-LTOP

[19] http://www.t9.com/

[20] http://www.ovum.com/go/product/IssuePaper/mms.htm

Mobile Processes and Mobile Channels

Kevin Chalmers
Napier University, Scotland

Imed Romdhami
Napier University, Scotland

Jon Kerridge
Napier University, Scotland

INTRODUCTION

Since the late 1990s there has been an explosion of new Internet technologies, making the Internet a primary medium for academic, business, and focus interest groups. As a result, new applications are constantly emerging and evolving. Couple this with the concept of mobility, something that is all around us today, and it can be seen that new applications require vastly different characteristics to those envisioned by the original Internet architects. For example, point-to-point IP traffic was designed to allow stationary parties to communicate, but when one of these parties becomes mobile with a wirelessly enabled mobile device, the original models become more difficult to implement.

This trend is a direct result of the fast development of mobile and wireless technologies, and it brings with it new problems. A nomadic user, equipped with a laptop or personal data device with wireless Internet connectivity, still requires and expects seamless communication while they migrate between locations. To overcome this and other issues, mobility is generally tackled at a hardware level, using and modifying existing underlying network techniques to accommodate this new mobile aspect. Designing a truly mobile system, however, also requires mobile software—software components that have the same freedom of movement as the hardware they are executed on.

BACKGROUND

Recent advances in programming frameworks, such as Microsoft's .NET framework and Sun Microsystems's Java platform, provide a number of mechanisms allowing applications to not only be portable between devices, but for applications to actually move from one device to another during runtime (Brooks, 2004; Delamaro & Picco, 2002). These types of systems fall under the general heading of code mobility platforms (Fuggetta, Picco, & Vigna, 1998), the same category as mobile agents and remote invocation systems. These interactions are usually modeled using more traditional techniques such as UML, but this does not truly allow us a very high-level view of how a mobile system actually behaves, being based on architectures aimed at much more static, single machine systems. Therefore a different paradigm should be employed, and here we examine mobile processes and mobile channels. Viewing a mobile component as a mobile process and a communication between them as a mobile channel, a much broader viewpoint of system mobility can be achieved, facilitating development, understanding, and maintainability.

Mobility

Mobility is all around us today. Wireless technology and mobile phone ubiquity have brought about a technological landscape that is no longer bound to a specific site, but can be carried globally from location to location moving around as we do. The trouble is we still expect to have the same level of service as we migrate as we get when we are stationary. This is of course difficult (imagine trying to hop from one moving train to another instead of having stations), and research into how seamless communication can occur is ongoing in such fields as mobile IP (Johnson & Perkins, 2004; Perkins, 2005) and migratory TCP (Sultan, Srinivasan, Iyer, & Iftode, 2002).

Here, we consider mobility more at the software level, although the basic concepts can be utilized for modeling lower level interactions. Indeed, most of the ideas presented are based on a more formal model—the π-calculus (Milner, 1999)—which has been used to model hardware processes such as mobile phone migration between base stations (cells). We do not, however, use this more formalized style here, but adopt a much softer view of mobility.

In the following paragraphs we will describe what mobile processes and channels are, and how they can be used to model some basic systems. These simple yet powerful building blocks are available in the JCSP (Communicating Sequential Processes for Java) package (Welch, Aldous, & Foster, 2002), which is based on yet another formal model, Communicating Sequential Processes (Hoare, 1978). Recent

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Symbols

Symbol D	escription
➞	A (stationary) channel
⬭	A (stationary) process
⇢	A mobile channel
◌	A mobile process

work (Chalmers & Kerridge, 2005) has simplified the usage of the mobility constructs within this architecture, allowing many interesting approaches to system development to occur, some of which shall be described presently.

SYMBOLS USED

Before moving on, we will define in Table 1 some basic symbols that shall be used.

SIMPLE PROCESS NETWORK

Processes communicate to each other using channels; synchronizing during a read-write operation (i.e., when a read occurs, the process waits until the associated write operation occurs, and vice versa). A set of processes interconnected by channels is generally referred to as a process network. Also of note is the fact that messages travel in one direction on a channel, as opposed to the two-way viewpoint normally thought of when considering network connections. For example if we consider a client-server application, using channels to model the connection, two separate channels will be created, one from the client to the server, and one from the server to the client. Using this level of interaction allows some interesting models to be built, especially related to events generated by messages being sent across a channel, which allows distributed event-based systems to be simply developed. Remember that channels synchronize on communication also, adding more interesting models.

Process and channel networks can be built up to a very high level of complexity if required, but individual processes themselves are generally quite simplistic with simple input and output interfaces. Processes can also contain other processes and channels operating within them, permitting the detail of a process network to be viewed at differing levels if required.

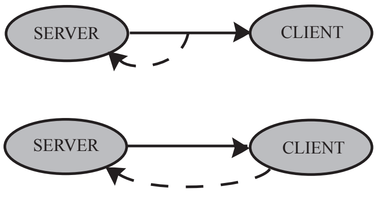
MOBILE CHANNELS

A mobile channel can be viewed in simple terms as a channel object that can be sent across a communication channel. This could be another channel altogether, or the channel input end could send its opposite output end if necessary (although the need for this is questionable). Milner (1997) considers this ability of communication a sign that it is sufficiently adult, in that it can utilize a property of itself. In other words the answer to the question *Can a communication communicate a method of communication?* is yes. Usually, only a single end (the input or output) is sent, allowing another process to communicate with the sender, or another process, over the received channel. If, for example, we examine a streaming server system, it is possible to create a client-server communication by the server sending a mobile channel end to the client, as in Figure 1.

The client process can now communicate back to the server using the mobile channel end. Mobile channels, just like normal channels, can also communicate channel ends. Or channels can be passed much further down a chain of connected processes, allowing remote systems to start communicating without prior knowledge of one another. This allows such network application interactions as peer-to-peer to be modeled. This would simply involve a peer requesting the output end for another peer from a central server, and the opposite doing the same for a connection back to it. This allows the two peers to communicate with each other without knowing where they specifically were before the interaction.

When we talk about a mobile channel end, we do not necessarily mean that the underlying software component is actually migrating around a distributed environment. This kind of interaction is possible, but it is also true to say that the location of the channel end can be passed and the necessary channel created dynamically when this location is received. If we consider channel mobility to also include this, then we can see that simply typing in the channel location (a URL for example) becomes a simulated mobile channel. In short, almost every distributed connection can be modeled as a mobile channel. What the more structured mobile channel enables is the movement of a connection without losing any sent or received messages, while still providing

Figure 1. Mobile channel



a synchronized message transfer handshake. This higher level abstraction is much more useful to the software practitioner, allowing channel movement to occur transparently in a single instruction. This is in contrast to the four-stage interaction of get location, send location, receive location, and create channel.

So, as can be seen, the ideas behind mobile channels have been present in systems for a long time; the viewpoint is just rarely taken, if at all. Modeling generally happens at an object level in most modern systems, but this tells us little about how distributed systems communicate or the structure of the communicating components. To illustrate how versatile this approach is, we will examine a much more common networking interaction, the three-way handshake.

Example: Three-Way Handshake

The three-way handshake occurs whenever a client-server communication is initiated within a TCP/IP environment. The three stages (hence the name) of the handshake are as follows:

1. The server creates a server socket to allow client connections.
2. The client connects to the server socket.
3. The server accepts the connection and obtains a new socket unique to that specific client-server communication.

By performing these three stages, a TCP connection provides the guaranteed transmission and reception of traffic that is a property of this protocol. Actually, we have already modeled this in Figure 1, although the three-way handshake is the reverse of this. The client has a connection from itself to the server this time, and the client sends a mobile channel end to the server to provide the unique connection. This interaction is currently being investigated in channel-based systems to allow multiple mobile clients to request services from a server when they enter the location of a wireless access point. As the server will not know about the devices until they request a service, and since each service requires its own access channel, then using mobile channels provides an easier level of interaction. The same interaction is also used to request channels from a server, an interaction that can be seen in the previous description of the peer-to-peer system previously discussed.

What people might argue here is that this does not accurately model a network connection, as a network stream generally allows two-way traffic. This is correct, but we view each channel as a one-way transmission method only. To provide a two-way method, a separate channel connection would be created. By viewing a connection as one-way only also provides a lot more control in a distributed environment, and multiple channel connections do not necessarily mean multiple sockets. For example, sockets are related to a different inter-node connection in JCSP, whereas channels are related to processes (nodes can be viewed as individual runtime environments or distributed machines). This allows multiple distributed processes to communicate back to a server using the same channel (an Any-2-One connection), even though they are using separate underlying sockets. This Any-2-One configuration can allow a number of interesting configurations, such as a formalized print spooler that can handle multiple users using multiple printers or a server that can handle multiple clients on the same input channel.

MOBILE PROCESSES

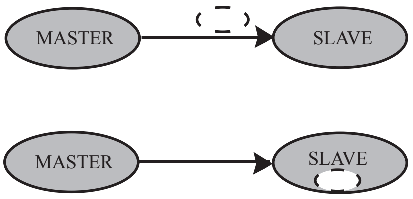
A simple way to view a mobile process is as a more state-conscious and controlled version of a mobile agent (White, 1994). Indeed, mobile agents and mobile processes tend to share many common properties, the ability to migrate from system to system being the one of prime importance. This requires some form of code mobility system that allows code accessible on one machine to be loaded onto another during runtime for execution by the receiving machine. A mobile process differs from an agent in that it can be much more sophisticated, possibly including internal processes or even being almost a full-scale application. Compare this to a mobile agent, which is usually a small task-orientated software component. That is not to say that a mobile agent is not a useful tool, but that it could not be used to model a mobile process, whereas a mobile process can model the mobile agent.

To allow a truly mobile process, it has to be possible for the receiving node to request and load the necessary data required to invoke the process. In JCSP for example, this is handled by dynamic loading of Java class data, but any form of class loading, or even scripting, could provide the same result. The advantage of using a compiled language, such as Java or a .NET derivative, is that the system will run faster than its scripting counterparts. This code loading can be hidden inside a channel (as is the case in JCSP), meaning that the developer does not have to consider it, merely accepting that if an object is sent between nodes then the class data will be loaded as required. As an example of how a mobile process would behave, we will look at how a simple task-orientated mobile agent is modeled.

Example: Task-Orientated Mobile Agent

The task-orientated agent design pattern (often called the master-slave agent design pattern (Ariador & Lange, 1998)) involves the sending of an agent from a master (or host) system to a slave (or client) system. Figure 2 illustrates this interaction.

Figure 2. Mobile process



All the master process simply does is connect to the slave and send the mobile process down the previously created channel. When the slave process receives the mobile process, it need only invoke it (by calling a method on the process to run it) and the slave will then execute the task. Again, this simple interaction can be built upon to provide much more complex systems, including mobile processes that can create and distribute mobile processes themselves. Any required class loading is handled within the underlying channels, providing a viewpoint that appears to simply be two distributed nodes passing an object between them. And this is basically what is happening in the underlying interaction.

COMBINING PROCESSES AND CHANNELS

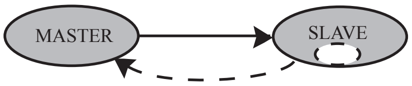
The previous example does not really provide an accurate model of how a mobile agent would behave in practice. Really what would happen is that the agent would return after performing its task, carrying with it some data relevant to the executed task. Or it may be in constant communication with the master process, providing feedback on its execution as required. This is easily modeled by combining both mobile processes and channels together, and therefore we will look at this to provide a means of modeling a more accurate mobile agent.

Example: Returning Agent

The model for a returning mobile agent is provided in Figure 3. As we can see, all that needs to be added is a mobile channel end within the mobile process that connects it back to the master process. Whenever the mobile process needs to communicate back to its master, it can do so using this channel end. The mobile process can also return to the master by writing a copy of itself back down the channel, then destroying the version left on the slave.

It is also possible to model a traveling mobile agent using these techniques, by having the mobile process carry multiple mobile channel ends within it, and using these to send it to different nodes as required. This allows the agent to migrate around a distributed environment, executing tasks at the required nodes.

Figure 3. Returning agent



FUTURE WORK

The ideas presented here are currently being used to develop location-enabled (site-specific) services to mobile devices by using connection migration and class loading techniques. The hope is that it will be possible to provide a service to a mobile device when it enters a location, without the device requiring any prior knowledge of the service. The service itself could be as simple as a text message that appears on entry into a wireless access point's coverage area, or as complex as allowing interaction with local hardware components (projector, printer, etc), providing location-specific games or site-specific services. These services are of particular interest, especially within the ubiquitous computing community (Toye, Sharp, Madhavapeddy, & Scott, 2005), as they are powerful yet easy to understand for the common user (an electronic ticket machine at a train station is a site-specific service). By allowing multiple users to use a service such as this via their mobile phones, we remove the need for queues, easing our everyday lives. All this can be done with the mobile device having minimal prerequisites, all necessary software being loaded dynamically as required then disposed of when finished. The hope is to provide an inroad into more universal ubiquitous interaction between devices, without forcing them to behave in a certain way.

An easy way to envision how this works is to consider how two people communicate in comparison to how two devices communicate. Humans have various methods of communication, such as writing and speech. However, if two people do not understand each other's language, then much more primitive methods of communication need to be used, such as hand signals. This continues until one or the other learns enough of their opposite's language to allow a higher level of communication.

Looking at mobile devices, they generally share some method of primitive communication via a wireless network interface. But they can rarely truly interact with one another, as they do not talk the same language. Yes, simple data can be passed, and XML can provide some structure to this; but to truly interact, one device must teach its language to another—and that can be viewed as one device passing a mobile process to another. By doing this, it would be possible for two devices to spontaneously interact, allowing, for example, two people to play a networked game with only one actually requiring the game on his or her device. Or it could spell the end to remote controls for each different device in the home. For example, a TV could send its remote control interface to a mobile phone, allowing the mobile phone to control the TV without any prior knowledge. In short, providing a framework that can load processes and create dynamic connections is a step towards a truly ubiquitous computing landscape, as envisioned by Weiser (1991).

SUMMARY

In summary, we have examined the interactions possible when viewing mobility at the level of channels and processes. This higher level view aids the understanding of how mobile systems interact, and aids our ability to design interacting systems such as those required for ubiquitous computing. Modeling systems in such a way allows mobile systems to be envisioned in simpler terms, taking the process view as opposed to the object view. Ideas from mobile agents to simple network interactions have been examined, providing a broad view as to what is possible when utilizing this viewpoint.

REFERENCES

Ariador, Y., & Lange, D. B. (1998). Agent design patterns: Elements of agent application design. *Proceedings of the 2nd International Conference on Autonomous Agents,* Minneapolis, MN.

Brooks, R. R. (2004). Mobile code paradigms and security issues. *IEEE Internet Computing, 8*(3), 54-59.

Chalmers, K., & Kerridge, J. (2005, September). jcsp.mobile: A package enabling mobile processes and channels. *Proceedings of the Communicating Process Architectures Conference 2005,* Eindhoven, The Netherlands.

Delamaro, M., & Picco, G. P. (2002, October). Mobile code in .NET: A porting experience. *Proceedings of the 6th International Conference Mobile Agents (MA 2002)*, Barcelona, Spain.

Fuggetta, A., Picco, G. P., & Vigna, G. (1998). Understanding code mobility. *IEEE Transactions on Software Engineering, 24*(5), 342-361.

Hoare, C. A. R. (1978). Communicating sequential processes. *Communications of the ACM, 21*(8), 666-677.

Johnson, D., & Perkins, C. (2004). *Mobility support in IPv6* (No. RFC 3775). IETF.

Milner, R. (1997). *Turing, computing and communication.* Retrieved March 10, 2006, from http://www.cl.cam.ac.uk/users/rm135/turing.pdf

Milner, R. (1999). *Communicating and mobile systems: The π-calculus.* Cambridge, UK: Cambridge University Press.

Perkins, C. (2005). *IP mobility support for IPv4* (rev. ed.). IETF.

Sultan, F., Srinivasan, K., Iyer, D., & Iftode, L. (2002). Migratory TCP: Connection migration for service continuity in the Internet. *Proceedings of the 22nd International Conference on Distributed Computing Systems.*

Toye, E., Sharp, R., Madhavapeddy, A., & Scott, D. (2005). Using smart phones to access site-specific services. *IEEE Pervasive Computing, 4*(2), 60-66.

Weiser, M. (1991, September). The computer for the 21st century. *Scientific American,* (September), 94-104.

Welch, P. H., Aldous, J. R., & Foster, J. (2002, April). CSP networking for Java (JCSP.net). *Proceedings of the International Conference on Computational Science (ICCS 2002)*, Amsterdam, The Netherlands.

White, J. (1994). *Mobile agents white paper.* General Magic.

KEY TERMS

Code Mobility: Code that can be dynamically loaded into another runtime environment during execution. This is essential for mobile processes and mobile agents.

Communicating Sequential Processes for Java (JCSP): A Java implementation of the formal model communicating sequential processes (CSP), which also incorporates features of the formal model the π-calculus.

Mobile Agent: A task-specific mobile process.

Mobile Channel: A communication construct that is able to be sent via another communication element.

Mobile Process: A process with the ability to migrate from one runtime environment to another.

Node: A runtime environment within a network. Nodes may exist on a single machine or may be distributed around a network.

Site-Specific Service: A service that is based in a particular location. In physical terms this could be considered as an electronic ticket machine.

Mobile Public Key Infrastructures

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A., Greece

George K. Lalopoulos
Hellenic Telecommunications Organization S.A., Greece

Stergios P. Chochliouros
Independent Consultant, Greece

Anastasia S. Spiliopoulou
Hellenic Telecommunications Organization S.A., Greece

INTRODUCTION

During the last decade we have witnessed the widespread penetration of mobile communications infrastructures and services (Chochliouros & Spiliopoulou-Chochliourou, 2005a) together with a great expansion of various information terminals such as cell phones, notebook computers, and personal digital assistants (PDAs). Broadband facilities (Chochliouros & Spiliopoulou-Chochliourou, 2005b) have also forwarded evolutionary processes and promoted the role of the mobile sector. In particular, improved performance of cell phones and their enhanced Web-based features have resulted in their use as personal trusted devices (PTDs), in order to perform tasks such as mobile banking, stock brokering, mobile ticketing, mobile shopping, access of corporate databases, and handling of e-government procedures (e.g., completing various types of documents) (May, 2001). These must be completed in a secure and safe environment that will guarantee protection from malicious or illegal actions like spoofing—namely, stealing information concerning financial transactions (such as passwords, bank account numbers, and credit card accounts), tampering with significant documents, and so forth.

From today's perspective, network and information security (European Commission, 2001) is about ensuring the availability of services and data; preventing the disruption and unauthorized interception of communications; confirming that data sent, received, or stored is complete and unchanged; securing data confidentiality; protecting information systems against unauthorized access; and protecting against attacks (involving malicious software and securing dependable authentication—that is, the confirming of an asserted identity of entities or users). Specific security measures therefore should be taken in order to establish an appropriate environment.

BACKGROUND

PKI Technology as the Means to Ensure Security

"Secure mobile transactions" implicates that, at least, the following specific features must be ensured satisfactorily:

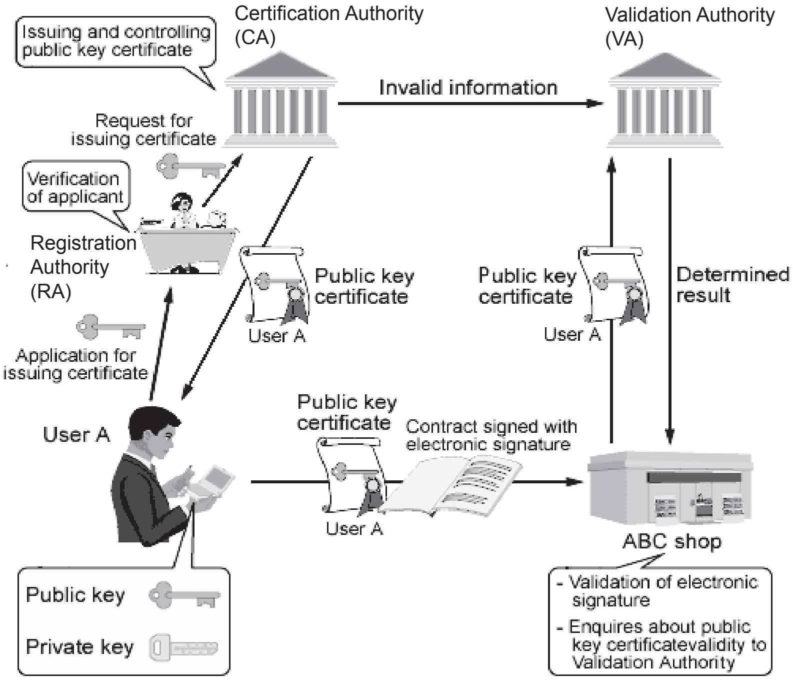
1. **Confidentiality:** The "assurance" that information is accessible only to those parties/entities that are appropriately authorized to access it.
2. **Integrity:** The assurance that information (either stored or transmitted) has not been altered (with or without intention) between two communication points or at a given time gap.
3. **Authentication:** The assurance that the source of information is "who" it claims to be.
4. **Non-Repudiation:** The assurance that communication parties remain honest about their actions—that is, that they cannot falsely deny having originated or received information.

Public key infrastructure (PKI) is able to offer (IETF, 2005) all the above security services for the case of mobile transactions. In the following sections, we provide a more detailed explanation of PKI.

PKI is based on a main tool for encrypting and decrypting data (Adams & Lloyd, 2003), which is called a "key." In fact, public key cryptosystems use two kinds of keys: a "public key" and a "private key." The latter is held by one (legal or physical) person and is for that person's unique use only; in contrast, the former is open to the public for widespread use. In the case of authentication by a public key cryptosystem, the "person"/"entity" subject to authentication starts by encrypting the transmitted data with his private key; resulting data

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. A fundamental PKI infrastructure


Issuing and controlling public key certificate
Certification Authority (CA)
Validation Authority (VA)
Invalid information
Request for issuing certificate
Verification of applicant
Registration Authority (RA)
Public key certificate
User A
Public key certificate
User A
Determined result
Application for issuing certificate
User A
Public key certificate
Contract signed with electronic signature
User A
ABC shop
Public key
Private key
- Validation of electronic signature
- Enquires about public key certificatevalidity to Validation Authority

cannot be read, unless a great deal of complex decryption is done; in fact, it cannot even be read by the person who encrypted it. Next, the "entity" realizing authentication uses the public key for data to decryption, and so information returns to a "readable" status. If data is correctly decrypted, the performer concludes that the key used for the encryption purposes was the private key that corresponds to the public key; consequently, the "person" who encrypted the data must be the holder of the private key. A fundamental question that raises here regards the case in which the entity performing the authentication mistakes the holder of the private key. Whether the encrypted transmitted data can be decrypted correctly simply depends on the specific nature of the public key corresponding with the private key. On the other hand, if the public key belongs to a complete "stranger" but does correspond to the proper private key, the stranger can decrypt the encrypted transmitted data. Therefore, authentication of a legitimate person can be mistaken, and it is possible that someone can pretend to be someone else.

The above-described scenario means that in the case of authentication through a public key cryptosystem, it is extremely important to correctly connect the right person and the public key. Consequently, it has become essential to devise a system that can certify, by means of utilizing a third-party organization with no direct connection to the person undergoing authentication, whether the person in question is unmistakably the person holding the corresponding private key or whether that person is a malicious "stranger" intending to spoof the cryptosystem. This scheme is called PKI, which practically constitutes a core technology that configures the security infrastructure for protecting electronic commerce (e-commerce), especially in the mobile sector (Lalopoulos, Chochliouros, & Spiliopoulou-Chochliourou, 2005).

A Social System for Supporting PKI Implementation

A commonly accepted configuration of a proper social system for the efficient support of PKI perspective is described (Kaji, 2004) in Figure 1.

Figure 1 shows the separate roles of various organizations, authorities, and other entities involved in authentication and certification procedures within a PKI system.

The first key-concept, a so-called certification authority (CA), confirms "who" is the owner of the private key corresponding to the public key and fixes the "prescribed" correspondence between them. According to existing regulatory provisions and practices (European Parliament and Council of the European Union, 2000), CAs can be public or private organizations. The CA then issues (and controls) an "electronic certificate" as the authorization of this correspondence. A registration authority (RA) is an organization responsible for verifying the identity of the key holder and checking his certification with the CA.

The second key-concept is the validation authority (VA), a body for checking the legality of electronic certifi-

Figure 2. The layered WAP architecture

Wireless Application Environment (WAE)
Wireless Session Protocol (WSP)
Wireless Transaction Protocol (WTP)
Wireless Transport Layer Security (WTLS)
Wireless Datagram Protocol (WDP)
Bearers (e.g., data, SMS [short message service], USSD [unstructured supplementary service data])

cates—namely, whether a certificate is valid and whether a trustworthy CA issued this. Since the PKI is a system to prevent spoofing, such a validity checking procedure is considered of basic importance among all operations.

The Concept of PKI in the Mobile Environment

From a technical point of view, mobile transactions over wireless networks are inherently insecure compared to electronic transactions over the Internet. The main reasons are:

- **Integrity:** Interference and fading make the wireless channel error-prone. Frequent handoffs and disconnections also degrade the security services offered.
- **Confidentiality:** The broadcast nature of the radio channel makes it easier to tap. Thus, communication can be intercepted and interpreted without difficulty if no security mechanisms (such as cryptographic encryption) are employed.
- **Authentication:** The mobility of wireless devices introduces an additional difficulty in identifying and authenticating mobile information terminals.

Therefore, specific measures are necessary to provide for enhanced security (European Commission, 2001), especially if considering the high penetration of mobile applications, both for corporate and residential users. On the other hand, there are several mobile transactions that do not require high security, like micro-payments and/or traditional Web-browsing. (For example, buying refreshments from a vending machine via the use of a cellular phone does not require an electronic signature.) The demand for strong authentication, however, becomes drastically higher in the case of mobile banking. One has therefore to make sure that his mobile banking transactions are safe, with no risk of a "third party" using his mobile phone illegally to "conduct" the activity.

PKI has proven itself a reliable solution for providing the necessary security frame. In fact, it can be used in mobile environments by security protocols in the same way as it is used in a fixed network. However, a specific adaptation of PKI is needed, so as to cope with the inherent limitations of mobile technology (e.g., processing power, memory size, communication bandwidth, and battery power). Two wireless communications solutions were developed in order to facilitate mobile transactions through Internet access: (a) the i-mode, and (b) the wireless application protocol (WAP). Since i-mode is a proprietary NTT DoCoMo scheme, our effort will focus on WAP (Open Mobile Alliance, 2002), which is a publicly available solution and a *de facto* standard for wireless communication.

THE WAP ENVIRONMENT

WAP is an industry-wide specification for developing applications that operate over wireless communication networks. It bridges the gap between the mobile and the Internet, as well as corporate intranets, and offers the ability to deliver a wide range of mobile value-added services to subscribers (information, location-based services, corporate information, interactive entertainment, etc.). The WAP specification defines an open, standard architecture and a set of relevant implementation protocols. WAP has a layered architecture (see Figure 2).

In the following paragraphs, we examine separately the features of each layer in the protocol stack.

Wireless Application Environment (WAE)

The WAE defines the user interface on the phone—that is, the application development environment to facilitate services supporting multiple bearers. To achieve this, the WAE contains: (1) a microbrowser for the wireless handset; (2) the Wireless Markup Language (WML), a presentation language for rendering WAP through the microbrowser; (3) WMLScript, a scripting micro-language similar to JavaScript; and (4) the wireless telephony application (WTA) that provides access to traditional telephony services (such as call-forwarding through WMLScript). These are the tools allowing WAP-based applications to be developed.

Wireless Session Protocol (WSP)

WSP is an intermediate layer that links the WAE to two session services: one connection-oriented operating above the wireless transaction protocol (WTP) and a connectionless service operating above the wireless datagram protocol (WDP).

Wireless Transaction Protocol (WTP)

WTP is a transaction layer providing transport services. It runs on top of a datagram service such as user datagram protocol (UDP); it is a part of the standard suite of TCP/IP (transport control protocol/Internet protocol) protocols, to provide a simplified one suitable for low-bandwidth mobile stations. Interestingly, WTP supports protocol data unit (PDU) concatenation and delayed acknowledgement to help reduce the number of messages sent. This protocol therefore tries to optimize the user experience by providing the necessary information, when it is needed. By stringing several messages together, the end user may well be able to get a better "sense" more quickly for what information is being communicated.

Wireless Transport Layer Security (WTLS)

The WTLS protocol (Open Mobile Alliance, 2002) is a PKI-enabled security protocol, designed for securing communications and transactions over wireless networks. It includes data integrity checks, privacy on the WAP gateway to client support, and authentication. It is used with the WAP transport protocols to provide security on the transport layer between the WAP client in the mobile device and the WAP server in the WAP gateway. Applications can selectively enable or disable WTLS services, depending on their security requirements and the characteristics of the underlying network. WTLS provides functionality similar to the Internet secure socket layer (SSL) and transport layer security (TLS) systems, but it has been optimized for use over narrow-band communication channels and incorporates datagram support. WTLS is being implemented in all major microbrowsers and WAP servers.

Wireless Datagram Protocol (WDP)

WDP is a connectionless transport layer allowing WAP to be bearer-independent (by adapting the transport layer of the underlying bearer). WDP presents a consistent data format to the higher layers of the WAP protocol stack, thereby conferring the advantage of bearer independence to application developers.

The WAP has evolved in various versions: WAP 1.0 → 1.1. → 1.2 → 1.3 → 2.0. The latest (WAP 2.0) release (Open Mobile Alliance, 2002) has one significant difference from the previous ones. In particular, WAP 1.x releases make use of a component called the WAP Gateway (or WAP Proxy) in order to translate Web-based protocols to/from WAP-based ones. However, the WAP 2.0 protocol does not require the WAP Gateway, since the mobile WAP 2.0 browser supports HTTP and standard Internet network and transport protocols (i.e., TCP/IP). By adding these protocols-standards and providing interoperable optimizations suitable to a wireless communications environment, the WAP specifications make available an appropriate framework that permits wireless devices to utilize existing Internet technologies. However, deploying a WAP Proxy in WAP 2.0 can optimize the communications process and may offer mobile service enhancements, such as location, privacy, presence- based services, and "push" functionality.

WAP 2.0 adopts the eXtensible HyperText Markup Language (XHTML) developed by the World Wide Web Consortium (W3C) as the language for rendering content, irrespective of whether it is created for the Internet or otherwise. XHTML provides a universal format for structured documents and data on the Web, while it has been extended to offer PKI-based functionality. A key management framework has been introduced, and standards for encryption and generation of signatures have been applied. The idea of key/certificate management in XML is to source out some of the functionality that is usually done by the client to a trusted party. This allows developers to build up only one version of content. WAP 2.0 is backwards compatible because WML (the language used by WAP 1.x) is a subset of XHTML. Some additional features of WAP 2.0 are:

1. **WAP Push:** This service allows content to be sent or "pushed" to devices by server-based applications via a "push" proxy. This capability is critical in delivering important information and alerts to the mobile devices without the user initiating a request.
2. **Provisioning:** This provides a standard approach to supply WAP clients with information needed to operate on the wireless network(s). It permits the network operator to use a common set of tools.
3. **MMS (Multimedia Messaging Service) Support:** WAP 2.0 has built-in relevant support, which allows MMS to seamlessly integrate with the wireless Internet interface. WAP 2.0 supports SyncML, the XML protocol that allows synchronization for all devices (particularly mobile).

VARIOUS ASPECTS OF THE WAP SECURITY

In this section we examine the security components of the WAP, recent developments, and possible future directions.

WTLS/TLS

WTLS/TLS is the basis of WAP 1.x security in the WTLS protocol, a transport-level security protocol (Grech, 2005). The WTLS specification allows for three classes (levels) of relevant implementation:

- **Class 1: Anonymous Encryption:** Data is encrypted, but certificates are not exchanged between the client and the gateway.
- **Class 2: Encryption with Server Authentication:** Data is encrypted and the client requires a digital

certificate from the server. Server-side authentication is performed using public key certificates similar to the SSL/TLS protocol. The WAP Gateway uses a WTLS certificate (a particular form of X.509 certificate compressed to save on bandwidth).
- **Class 3: Encryption with Client and Server Authentication:** Data is encrypted and the client and the server exchange digital certificates. Clients are able to authenticate, using proper certificates. These are of regular X.509 format and can be stored either on the client or on a publicly accessible server (where a pointer to the certificate will be stored on the mobile device).

WAP 2.0 uses TLS instead of WTLS due to requiring end-to-end security with all-IP-based technology in order to overcome the WAP gateway security breaches. For example, sensitive information can be translated into clear texts, so the operator may read it at the gateway (the "WAP gap"). A possible solution is to make the WAP Gateway resident within an enterprise (server) network, where heavyweight security mechanisms can be enforced. The WAP 2.0 overcomes this problem by using TLS tunneling to support end-to-end security at the transport level. TLS is a PKI-enabled protocol that provides services such as authentication (by using digital signatures and public key certificates), confidentiality (by encrypting wireless data), integrity (by employing hashing of wireless data for detecting data modifications), and denial-of-service protection (by detecting and rejecting data that has been replayed or not successfully verified).

WMLSCrypt

WML script crypto (WMLSCrypt) is an application programming interface (API) that allows access to basic security functions (Ashley, Hinton, & Vandenwauver, 2001) in the WML script crypto library (WMLSCLib), such as key-pair generation, digital signatures, and the functions that process objects commonly found in the PKI (e.g., keys and public-key certificates). WMLSCrypt allows WAP applications to access and use the security objects and basic security services managed by other WAP security standards. WML Script can utilize an underlying WIM Module to provide the crypto-functionality.

WAP Identity Module (WIM)

The WIM is a tamper-resistant computer chip that optionally resides in WAP-enabled devices (such as mobile phones and PTDs). It can store key material like the PKI root public key and the user's private key. WIMs are most commonly implemented using smart card (Chochliouros & Spiliopoulou-Chochliourou, 2002) chips that have memory and storage for data and programs. A secure solution is to integrate WIM functionality into a SIM (subscriber identity module) card, resulting to the SWIM (Subscriber WAP identity module) card.

The SIM card is by itself a smart card, and the reader for this is already in place. It works as follows: When a mobile user wants to sign a transaction, a SMS with a signing request is sent to his phone from the institution he has business with (e.g., public department). To proceed, he only enters the signing PIN (personal identification number) code that activates the private keys and the signature function for the received SMS. The returned signature SMS is verified by the business by retrieving the certificate. If needed, transactions can be time-stamped and stored into the database, accessible by all parties involved in the procedure (i.e., both users and merchants), to assure the non-repudiation feature. Similar functions are available for mobile Web browsing, as the same signature function can be used when logging into some protected Web page. Such solutions are already deployed in Scandinavian countries (like Norway and Finland). Furthermore, it seems a preferred authentication/signature method in the wider European Union (EU) territory.

Wireless Public Key Infrastructure (WPKI)

WPKI is not an entirely new set of standards for PKI; it is more an "optimized" extension of traditional PKI for the wireless environment. Both WPKIs and PKIs implement mobile commerce businesses by managing relationships, keys, and certificates. WPKI is concerned primarily with the policies that are used to manage electronic businesses and the security environment by WTLS/TLS and WMLSCrypt. In the case of wired networks, IETF PKI standards are the most commonly used; for wireless networks, WAP's Forum WPKI standards are the most widely used.

WPKI requires (Yeun & Farnham, 2001) the same components as a traditional PKI—that is, an end-entity application, a RA, a CA, and a PKI repository. However, in WPKI, the end-entities and the registration authority are implemented differently, using the PKI portal. The latter is a network server, like the WAP Proxy. It typically functions as the RA and is responsible for translating requests made by the WAP client to the RA and CA. (It interoperates with WAP devices on the wireless network and with the CA on the wired network). The end-device application in WPKI is implemented as optimized software (based on WMLSCrypt API) that runs in the WAP device and performs the same functions as in traditional PKI. Moreover, it includes relevant functionalities. WPKI has optimized the PKI protocols, the certificate format, and the cryptographic algorithms and keys with respect to mobile environments. Compared to a PKI, WPKI applications have to work in an environment with less powerful CPUs (central processing units), less memory, restricted power consumption, smaller displays, and diverse input devices. Despite these shortcomings, the wireless equipment must be able to generate and register keys; manage end user mobile identities; encrypt and decrypt messages; and receive, verify, store, and send certificates/digitally signed data. In many cases, ordinary PTDs are not able to fulfill all these requirements, as they do not possess sufficient memory. Sometimes, a client's functionality has to be implemented outside the mobile equipment. Therefore, some WPKI solutions are very likely to employ "network agents" to perform some of these tasks. The PTDs must at least be able to perform a digital signature function to permit the establishment of a WPKI. Network agents can perform all other PKI-related tasks such as validation, archiving, or certificate delivery—that is, an implementation in which private keys are stored in a proxy server or alternatively, embedded into the tamper-resistant modules (like WIM/SWIM) of PTDs. Unfortunately, such solutions require more improvements, particularly in the area of key-pairs generation by end users, rather than being assigned by the network operators. In addition, a lack of standardization presents a major barrier in the development of wireless PKI. In other words, establishing trust in a WPKI is crucial for the success of applications that will exploit probable opportunities created by PTDs. Trust is based on the reliability of the technology, but also on a carefully implemented system of laws, policies, standards, and procedures (Chochliouros & Spiliopoulou-Chochliourou, 2003), which includes the management of certificates by trusted certificate authorities. The questions of anonymity, privacy, government surveillance, and industry-based policies (and standards) represent challenges that all parties involved have to face, if they plan to strengthen the level of trust that recent legislation efforts have already made possible throughout the world.

CONCLUSION

The high penetration of wireless devices paves the way for mobile commerce. Such devices have begun to transform into PTDs, suitable for applications like mobile banking, mobile shopping, citizen services, and so forth. A secure mobile infrastructure needs to ensure that such transactions are confidential, that all parties involved are clearly identified, and that the relevant agreements are non-reputable. This establishes a reliable and convenient framework for valid contracts, signed with digital signatures. Consequently, PTDs will be able to generate legally binding digital signatures and to enable a user to authenticate him remotely over the underlying networks. Such moves will quickly make mobile commerce a part of our everyday life. In particular, businesses are expected to extend their services to customers on the move.

To this aim, PKI can play an important role in meeting the corresponding security requirements. In a mobile environment, a special type of PKI (called WPKI) is used, in order to deal with the inherent limitations of wireless devices. WPKI includes most of the concepts that are present in traditional PKI. The combination of WPKI and the widely adopted standard WAP can offer users a platform for trusted and secure services, and significant new revenue opportunities for authentication, payment, and validation services.

In the near future, mobile operators face the challenge of providing secure authentication and value-added services between the PTDs and service/content providers. Their task will be to perform fundamental functions such as encryption/decryption, certificate validation, and key generation. Moreover, manufacturers also face the challenge of making wireless devices that are compact, powerful, easy to-use, and with single log-on security mechanisms. Each of these factors will contribute to the success of WPKI, and companies that are to develop, produce, and sell these products inexpensively (with the required quality), and will play an important role to gain an extremely large market.

Collaborative projects (WPKI Non-Profit Association, 2005) can perform prime actions towards developing a framework for stimulating the mobile market and establishing the appropriate environment for WPKI security promotion. Some significant factors for the development of WPKI are issues regarding WIM/SWIM cards, in combination with 3G technology, and the management of certificates via the CAs. To this aim, root certificate authority services and certificate chains are being used in the verification procedure (Developer Connection, 2005). Establishing a WPKI will generate potential for additional services (like those required for managing the financial risk involved in certification practices). Another major area of potential business can be found in offering and packaging a variety of certificate-related services like directory services, a notary service, or services for key generation or archiving. Others will be involved in monitoring technical compliance with policies and regulations.

The future challenge is to establish adequate trust in WPKI as a "new" medium, while the success of wireless applications will depend on their usefulness. The easy-to-use solutions are more likely to succeed than complicated ones. Governments, businesses, and end users can benefit from this. In order to enable secure mobile commerce solutions, qualified legal expertise needs to be considered in every environment (European Commission, 2004), together with the continuous enhancement of the WPKI standards throughout the world (Funk, 2003; Costello, 2002).

REFERENCES

Adams, C., & Lloyd, S. (Eds.). (2003). *Understanding PKI: Concepts, standards, and deployment considerations.* Boston: Addison-Wesley.

Ashley, P., Hinton, H., & Vandenwauver, M. (2001). Wired versus wireless security: The Internet, WAP and i-mode for e-commerce. *Proceedings of the 17th Annual Computer Security Application Conference (ACSAC)*. Retrieved October 28, 2005, from http://www.portal.acm.org/citation.cfm?id=872163

Chochliouros, I. P., & Spiliopoulou-Chochliourou, A.S. (2002). Smart cards—An overview of current European initiatives to ensure efficient deployment within the context of information society applications. *Proceedings of the EURESCOM Summit 2002—Powerful Networks for Profitable Services* (pp.167-174). Heidelberg, Germany: EURESCOM.

Chochliouros, I. P., & Spiliopoulou-Chochliourou, A. S. (2003). Innovative horizons for Europe: The new European telecom framework for the development of modern electronic networks and services. *Journal of the Communications Network, 2*(4), 53-62.

Chochliouros, I. P., & Spiliopoulou-Chochliourou, A. S. (2005a). Visions for the completion of the European successful migration to 3G systems and services—Current and future options for technology evolution, business opportunities, market development and regulatory challenges. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities* (pp. 342-368). Hershey, PA: IRM Press.

Chochliouros, I. P., & Spiliopoulou-Chochliourou, A. S. (2005b). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *International Journal of Infonomics, 1,* 5-21. London: E-Centre for Infonomics.

Costello, D. (2002). *Preparing for the m-commerce revolution—Mobile payments* (white paper). Dublin, Ireland: Trintech. Retrieved November 28, 2005, from www.aecomo.org/kbase/library/documents/mpayment_paper.pdf

Developer Connection. (2005). *Digital certificates.* Retrieved December 20, 2005, from http://developer.apple.com/documentation/Security/Conceptual/Security_Overview/Concepts/chapter_3_section_7.html

European Commission. (2001, June 6). *Communication on network and information security: Proposal for a European policy approach* (COM(2001) 298 final). Brussels, Belgium: European Commission.

European Commission. (2004). *Application of the e-money directive to mobile operators—Consultation paper of DG Internal Market*. Brussels, Belgium: Directorate General Internal Market of the European Commission. Retrieved December 12, 2005, from http://europa.eu.int/comm/internal_market/bank/docs/e-money/2004-05-consultation_en.pdf

European Parliament and Council of the European Union. (2000). *Directive 1999/93/EC on a community framework for electronic signatures. Official Journal L13,* (January 19), 12-20. Brussels, Belgium: European Parliament and Council of the European Union.

Funk, J. (2003). *Mobile disruption: The technologies and applications that are driving the mobile Internet.* Hoboken, NJ: John Wiley & Sons.

Grech, S. (2005). *Wireless security tutorial.* Lancaster, UK: Lancaster University, Management School. Retrieved January 7, 2006, from http://www.lancs.ac.uk/postgrad/grech/wsecurity.htm

IETF (Internet Engineering Task Force). (2005). *Public Key Infrastructure (X.509) pkix.* Retrieved, January 15, 2006 from http://www.ietf.org.html.charters/pkix-charter.html

Kaji, T. (2004). *PKI framework for supporting the security of mobile communication from its core.* Systems Development Laboratory, Hitachi Ltd. Retrieved December 20, 2005, from http://www.sdl.Hitachi.co.jp/English/topics/pki

Lalopoulos, G. K., Chochliouros, I. P., & Spiliopoulou-Chochliourou, A. S. (2005). Evolution of mobile commerce applications. In M. Pagani (Ed.), *The encyclopedia of multimedia technology and networking* (pp. 295-301). Hershey, PA: IRM Press.

May, P. (Ed.). (2001) *Mobile commerce: Opportunities, applications, and technologies of wireless business.* Cambridge, UK: Cambridge University Press.

Open Mobile Alliance. (2002). *WAP 2.0 conformance release.* Retrieved November 20, 2005, from http://www.openmobilealliance.org/tech/affiliates/wao/wapindex.html

Yeun, C., & Farnham, T. (2001). *Secure m-commerce with WPKI.* Lecture Notes in Computer Science, Toshiba Telecommunication Research Laboratory, Toshiba Research Europe Limited, Bristol, England. Retrieved December 12, 2005, from http://www.iris.re.kr/iwap01/program/download/g07_paper.pdf

WPKI Non-Profit Association. (2005). *WPKI project and infrastructure.* Retrieved October 10, 2005, from http://home.swipnet.se/susss/wpki/filerna/WPKI_general_presentation_1.pdf

KEY TERMS

Certification Authority (CA): An entity that issues/updates/revokes public-key bearing certificates in response to authenticated requests from legitimate registration authorities.

Public Key Infrastructure (PKI): Allows users to exchange data securely over an unsure network (such as the Internet), and involves the use of a public and private cryptographic key pair, which is obtained through a *trusted authority.* PKI provides for digital certificates that can identify individuals or organizations, and directory services that can store and revoke them.

Registration Authority (RA): An entity authorized to make requests to issue/revoke/update certificates to a CA. The registration authority can be considered similar to an account manager in function and is responsible for member enrollment and/or attribute assignments.

Trusted CA Information: Information that is stored by a PKI entity and indicates that a given certification authority is trusted as a root CA by that entity. This has to include a public key, which typically also includes a name and a validity period (often stored in the form of a self-signed certificate).

WAP Identity Module (WIM): Used in performing WTLS and application-level security functions, and especially to store and process information needed for user identification and authentication. WIM may be used by WPKI for secure storage of certificates and keys.

Wireless Application Protocol (WAP): A secure specification allowing users to access information instantly via handheld wireless devices such as mobile phones, pagers, two-way radios, smart phones, and communicators. WAP is a widely used set of protocols that standardize the manner in which wireless devices are able to access parts of the Internet (e.g., e-mail and the Web).

Wireless Public Key Infrastructure (WPKI): An optimized extension of traditional PKI for the wireless environment. Both WPKIs and PKIs support mobile commerce by managing relationships, keys, and certificates. WPKI is concerned primarily with the policies used to manage electronic businesses and security environment.

Wireless Transport Layer Security (WTLS): A security layer of the WAP, providing privacy, data integrity, and authentication for WAP services. It is needed because the client and the server must be authenticated in order for wireless transactions to remain secure and because the connection needs to be encrypted. WTLS is based on the widely used TLS v1.0 security layer used on the Internet.

Mobile Serverless Video Communication

Hans L. Cycon
FHTW Berlin, Germany

Thomas C. Schmidt
HAW Hamburg, Germany

Matthias Wählisch
FHTW Berlin, Germany

INTRODUCTION

Voice and video conferencing have been well established as regular communication services within the wired Internet. Facing the paradigm of ubiquitous computing and mobile communication, they are on the spot to be launched within a wireless Internet infrastructure. Following an 802.11, 802.16 or 3G standard, wireless networks provide enough bandwidth to support data intensive communication services such as videoconferencing. The vision of nomadic users at roaming devices performing synchronous communication, such as voice or videoconferencing over IP (VoIP/VCoIP), is around, but raises new challenges for the Internet infrastructure.

In conferencing scenarios addressability raises the first major issues. To globally call a device, a routable IP address must be in use. On a large scale such address space is only provided by IPv6. To identify a communication partner's current device, a supplementary global user locating scheme is needed. In wireless infrastructures, where users share limited bandwidth from a restricted frequency space, multicasting is needed to enable group conferencing compliant to transmission resources and without placing the burden of dedicated group-server infrastructure.

At the same time, synchronous real-time applications, such as VoIP and VCoIP place new demands on the quality of IP mobility services: packet loss, delay and delay variation (jitter) in a constant bit rate scenario need careful simultaneous control. A spoken syllable is about the payload of 100 ms continuous voice traffic. Each individual occurrence of packet loss above 1%, latencies over 100 ms or jitter exceeding 50 ms will clearly alienate or even distract the user. Audio and visual streams in video conferencing additionally require tight synchronization. Inter-stream latencies should remain below 30 ms for audio arriving ahead, 40 ms for audio being behind. While uni-directional distribution may compensate quality deficits by buffers, available techniques of hiding packet loss at the cost of delay and jitter or vice versa are of limited use in conferencing. Their requirements impose strong challenges on a mobile Internet scenario. Challenges are even tightened by multicast-based group communication, since in conferencing scenarios each member commonly operates as receiver and as sender. Real-time requirements consequently are a major driving force for the development of a seamless mobile Internet layer.

In concordance with communication capabilities, video coding techniques have evolved, as well. The latest standard for video coding H.264/AVC (ITU H.264, 2005), although designed as a generic standard, is predestined for applications like mobile video communications (Stockhammer, Hannuksela, & Wiegand, 2003). Besides enhanced compression efficiency, it delivers also network friendly video representation for interactive (video telephony) and non-interactive applications (broadcast, streaming, storage, and video on demand). H.264/AVC provides gains in compression efficiency of up to 50% over a wide range of bit rates and video resolutions compared to previous standards. While H.264/AVC decoding software has been successfully deployed on handhelds, high computational complexity still prevents pure software encoders in current mobile systems. There are, however, also fast hardware implementations available (see a list in Wikipedia, H.264, 2006). Next generation codes, like scalable video coding (SVC) are already in a design state (Reichel, Schwarz, & Wien, 2005; Schwarz et al., 2004). The main new feature, scalability, addresses schemes for delivery of video to diverse clients over heterogeneous networks, particularly in scenarios where the downstream conditions are not known in advance. The basic idea is that *one* encoded stream can serve networks with varying bandwidths or clients with different display resolutions or systems with different storage resources, which is an obvious advantage in heterogeneous networks prevalent in mobile applications.

BACKGROUND

Video conference communication is a person-oriented, session-based service. A caller requesting contact to one or several partners will expect to address a personal identifier, but establish the corresponding conference session with the devices currently in use by the callees. Unlike in mobile

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

telephony, the Internet architecture is required to locate users and mask the user-device mapping, following the paradigm of location transparency, like in e-mail services. Once established, sessions need to persist while mobile devices roam. Intermediate handovers thereby should unnoticeably comply with quality of service measures for real-time communication. Operating on portable devices with limited capacities of CPU, batteries and displays, a video conferencing solution needs to balance out network efficiency and adaptability versus coding complexity. Lightweight flexible software systems as introduced by Cycon et al. (2004) are preferably employed in mobile communication.

Conference Signaling

The traditional, ISDN compatible architecture of VCoIP systems has been defined in the standard ITU H.323 (2000). Central parts of this model are derived from a client-server principle with a Gatekeeper, providing connection control and address translation, and a multipoint control unit (MCU) serving video streams in multipoint conferences. The H.323 architecture must be considered local and immobile in the sense that all participants need to agree on common MCU and pre-configured Gatekeeper servers, which, at least for the MCU, suffer from severe scaling deficiencies.

A flexible, fairly general Internet signaling solution has been presented with the session initialization protocol (SIP) (Rosenberg et al., 2002). Beside user localization, SIP covers negotiations about user capabilities, user availability, the call set-up by session description protocol (SDP) data and the handling of the calls itself. SIP introduces its own infrastructure of servers, which actively perform a peer-to-peer routing by using SIP-URLs. SIP is based on an extensible method framework and open to store persistent data. SIP liberates the rigid addressing scheme of telephone numbers used in H.323, proposing addresses of the "e-mail-like" form <user>@<SIP-server>. A basic interaction with IP multicast is defined in SIP through the *maddr* address attribute in the VIA header.

Mobility Management

As an application layer protocol SIP provides some mobility support to session-based services (Wedlund & Schulzrinne, 1999), which requires implementation at the application layer. Employing the regional SIP server as an application specific home agent, handoff notifications are traded via regular SIP messages to the home server (register) and the correspondent node (re-invite). As SIP mobility operates above the transport layer, it remains self-consistent and transparent to the Internet infrastructure, but inherits all underlying delays in addition to its own signaling efforts.

The fundamental approach to mobility management in the next generation Internet is the Mobile IPv6 (MIPv6) RFC (Johnson, Perkins, & Arkko, 2004). MIPv6 transparently operates address changes on the IP layer as a device moves from one network to the other by sustaining original IP addresses in a home address destination option and hiding the different routes to the socket layer. In this way, hosts are enabled to maintain transport and session connections when they change locations. An additional infrastructure component, the MIPv6 Home Agent, preserves global addressability, while the mobile node is away from home.

Local handovers in MIPv6 are rapidly completed. In the presence of layer 2, triggers for movement detection, the time needed for address reconfiguration and local updates remains well below 10 ms (Schmidt & Wählisch, 2003). Distributed Mobile IPv6 scenarios, though, inherit strong topology dependence from binding updates with the Home Agent and correspondent nodes. To resolve topologically originated delays, Hierarchical MIPv6 (Soliman et al., 2005) has been introduced for micro mobility scenarios and Fast Handovers (Koodli, 2005) for delay hiding by means of handover predictions. Even though an expected disappearance of predictive handover delays does not hold in practice, these accelerating schemes arrive at real-time compliant handover performance (Schmidt & Wählisch, 2005).

IP layer handovers thus can be considered capable of a mobility management for real-time voice and video communication. SIP application layer handovers have been found by Kwon, Gerla, Das, & Das (2002)) to significantly fall behind Mobile IPv4, which itself is largely outperformed by MIPv6. Handoff disruptions of the underlying layer 2 add to service degradation, admitting vendor specific, but large values in 802.11 systems (Mischra, Shin, & Arbaugh, 2003). Further quality of service issues of wireless transmission technologies and of the general IP routing layer need consideration, as well.

Video Coding

H.264/AVC, or MPEG-4 Part 10, (formally, ISO/IEC 14496-10) is a digital video codec standard. It was defined by a collective effort of the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Video Coding Experts Group (VCEG) and the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) Moving Picture Experts Group (MPEG) known as the Joint Video Team (JVT). The final drafting work on the first version of the standard was completed in May of 2003.

H.264 is a name related to the ITU-T line of H.26x video standards, while AVC (advanced video coding) relates to the ISO/IEC MPEG side of the partnership project that completed the work on the standard. H.264/AVC has also been referred to as "the JVT codec". The intent of H.264/AVC project has been to create a standard that would be capable of providing good video quality at bit rates that are substantially lower than what previous standards would need (e.g., relative to MPEG-2, H.263, or MPEG-4), and to do so without much of an increase in complexity (Wiegand, Sullivan, Bjøntegaard, & Luthra, 2003; Ostermann et al., 2004). An additional goal was to work well on a very wide variety of networks and systems (e.g., mobile network systems).

H.246/AVC, as almost all conventional video codecs which are in use today, is not in a simple, efficient, application-friendly way scalable, that is, it encodes video content using a fix bit-rate tailored to a specific application. Video streaming in heterogenic networks today is realized by parallel encoding the video data into different compression/quality stages depending on the available bandwidth of the client (multibit streaming, sure stream, etc.). The disadvantages of those methods are obvious: unreasonable large memory and processing power is needed. Current digital video applications require at least three types of scalability features: quality scalability, spatial resolution scalability, and temporal (frame rate) scalability. Next generation codecs like scalable video coding (SVC) promise to solve this problem. The Moving Picture Experts Group (MPEG) started to work on (SVC) in December 2001. The goal was to provide scalability at the bit stream level, with good compression efficiency, allowing free combinations of scalable modes (such as spatial, temporal and SNR/fidelity scalability). In October 2003, MPEG issued a "Call for Proposals on Scalable Video Coding Technology". It was widely believed that the so-called 3-D wavelet coding will solve the problem efficiently (Choi & Woods, 1999; Ohm, 2002; Flierl, 2003). But it turned out that extensions of H.264/MPEG-4 AVC outperformed all other proposals and subjective tests for different scalability scenarios verified their superior coding efficiency (Wiegand et al., 2006). In January 2005, MPEG and the Video Coding Experts Group (VCEG) of the ITU-T agreed to jointly finalize the SVC project as an amendment of their H.264/MPEG-4 AVC standard, and the scalable coding scheme (Schierl, Schwarz, Marpe, & Wiegand, 2005; Schäfer et al., 2005; Schwarz, Marpe, & Wiegand, 2004).

MOBILE CONFERENCING SUPPORT

Locating Nomadic Users

A video conference is a synchronous form of communication, requiring online presence of its participants. These commonly register at some SIP server. The device a user is currently located at then can be easily identified by searching this SIP server's session directory. Locating a nomadic user therefore can be considered equivalent to identifying its SIP or session directory server. It can be easily achieved when addresses of the form <user>@<SIP-server> are employed. However, it is desirable and has been foreseen by SIP to allow for virtual addresses, which do not reflect existing machine infrastructure. Virtualization is not only demanded when telephone numbers are used, but has proven relevant within the global e-mail system.

In contrast to e-mail, though, conference sessions services are not built onto a DNS infrastructure like the MX record, which translates any virtual name into its e-mail server on duty. To enable virtual session naming services, Rosenberg and Schulzrinne (2002) propose to rely on DNS extensions, the SRV and NAPTR record, which are capable of encoding a service-to-server name mapping. This approach—even though technically irreproachable—suffers from a limited pervasiveness of DNS extension records. In practice, the allocation of DNS SRV records is very rarely seen. To overcome these obstacles, Schmidt, Wählisch, Cycon, and Palkow (2003) propose to restrict call-names to e-mail addresses and take advantaged of the globally established MX server record infrastructure by applying a name convention to session servers. This allows for a rollout of session server infrastructure in concordance with e-mail services. In proceeding along this line, session-oriented service support for nomadic users can be easily established, while Internet infrastructure remains unchanged.

Serverless Group Communication

An important application scenario for video communication is given by multi-party conferences. Ranging from a small group of few partners up too very large conventions, video conference technology faces the need to provide seamless, mobility transparent services to all of these assignments. The simple, server-centric approach of H.323 to group communication thereby is easily ruled out by scalability and flexibility issues.

Instead, as a complex and data intensive application video group conferencing requires relying on multicast distribution techniques, which have to account for media distribution as well as inter-party signaling. While the RTP architecture initially provides flexible multicast support for media streaming, a call control for multi-party usage of SIP is currently under development (Mahy et. al., 2005). Major focus is donated to unicast-based overlay multicast, introducing the function of "mixing node". Such mixers arrange group management and signal distribution substituting stationary servers and thereby attain a role incompatible with mobility constraints.

The basic SIP RFC 3261 already defines a message exchange using IP layer multicast: a client wishing to initiate or joint into a multi-party conference sends its INVITE request to a multicast group by employing the *maddr* attribute in the SIP VIA header. Group members subsequently indicate their presence by responding to the same group. Suitable for large, loosely coupled and mutually unknown parties, this simple scheme only operates through any source multicast (ASM). Due to its routing complexity and security threads, ASM deployment mostly remains restricted to single management domains. It is the general belief that a large-scale inter-domain deployment will be reserved to the emerging source specific multicast (SSM) (see mobile multicast).

To enable group communication by source specific multicast, SIP dialogs need alteration in the following way: a new conferencing member willing to join a previously established group conference invites any party and receives response including multicast session descriptions via unicast. In parallel, the invited party has to repeat its response to the previously established SSM signalling domain, in order to trigger an active source subscription with respect to the newly established caller at all previous group members. All additional group members subsequently will advertise their session affiliation, while the initially called party will signal a turnover of its newly established SIP signalling channel to SSM multicast. As soon as the new conferencing member has completed its subscription to SIP signalling and media sessions for all conference party's addresses, a source specific multicast group conference is fully established among peer-to-peer members in the absence of any coordinating instance (Schmidt, Wählisch, Cycon, & Palkow, 2006).

VIDEO CODECS FOR MOBILE SYSTEMS

H.264/AVC or MPEG-4 Part 10 contains a number of new features that allow it to compress video much more effectively than older standards and to provide more flexibility for application to a wide variety of network environments. In particular, such key features include (Richardson, 2006):

- Multi-picture motion compensation using previously encoded pictures as references in a much more flexible way than in past standards, thus allowing up to 32 reference pictures to be used in some cases.
- Variable block-size motion compensation with block sizes as large as 16 × 16 and as small as 4 × 4, enabling very precise segmentation of moving regions.
- Quarter-pixel precision for motion compensation, enabling very precise description of the displacements of moving areas.
- An in-loop deblocking filter, which helps to prevent the blocking artifacts common to other DCT-based image compression techniques.
- An exact-match integer 4×4 spatial block transform (similar to the well-known DCT design).
- Context-adaptive binary arithmetic coding (CABAC), which is a clever technique to lossless compression of syntax elements in the video stream, knowing the probabilities of syntax elements in a given context (Marpe, Schwarz, & Wiegand, 2003).
- Context-adaptive variable-length coding (CAVLC), which is a lower-complexity alternative to CABAC for the coding of quantized transform coefficient values.
- A network abstraction layer (NAL) definition allowing the same video syntax to be used in many network environments that provide more robustness and flexibility than provided in prior designs.

These techniques, along with several others, help H.264/AVC to perform significantly better than any prior standard can, it often performs radically better than MPEG-2 video—typically obtaining the same quality at half of the bit rate or less (Compression Links, 2006).

The standard includes six sets of capabilities, which are referred to as *profiles*, targeting specific classes of applications and some fidelity range extensions (Marpe, Wiegand, & Sullivan, 2005). We mention only three of them, which are relevant for mobile applications:

- **Baseline Profile (BP):** Primarily for lower-cost applications demanding less computing resources, this profile is used widely in videoconferencing and mobile applications.
- **Main Profile (MP):** Originally intended as the mainstream consumer profile for broadcast and storage applications.
- **Extended Profile (XP):** Intended as the streaming video profile, this profile has relatively high compression capability and increased robustness to data losses.

The next coding generation, scalable video coding (SVC) is defined as a codec architecture which is based on a layered representation (Wiegand et al., 2006, Schwarz, Marpe, & Wiegand, 2005, 2006). The design builds upon an H.264/AVC-compatible base layer, and re-uses existing elements such as motion-compensated prediction, intra prediction, transform coding, entropy coding, and deblocking filter while only a few components have been added or modified. Different scalable modes (such as spatial, temporal and SNR/fidelity scalability) can be combined to obtain much more flexibly as compared to scalable tools of previous standards.

Key features are:

- hierarchical prediction structure;
- layered coding scheme with switchable inter-layer prediction mechanisms;

- fine granular quality scalability using progressive refinement slices;
- usage and extension of the NAL unit concept of H.264 / MPEG-4 AVC.

FUTURE TRENDS

Infotainment has been well established within the mobile world. Digital radio telephony and (unidirectional) digital video broadcasting still govern the majority of mobile devices today, but emerging standards of a next-generation mobile Internet envision a network layer, which unifies hardware abstraction and offers scalable and flexible services at decreasing costs. Mobile video communication will be one foreseeable service of ubiquitously available Internet devices, superseding today's static multimedia messaging. Communication protocols and standards are rapidly evolving towards this vision.

Video encoding capabilities still are a major barrier today. SVC, though, as defined for the new generation standard based on the H.264/AVC will get in more efficient and less complexity versions hardware and also software implementations. This will be the basis software for a large variety of mobile video-capable communication gear. Application fields will be video streaming over heterogeneous IP networks, surveillance systems, mobile streaming video, wireless LAN video, multi-channel video production and distribution, layered protection of content, multi-party video telephony/conferencing and wireless broadcasting (Ohm, 2005).

CONCLUSION

Multimedia communication is a challenge to both the stationary and the mobile Internet; video conferencing simultaneously raises issues of system, network and protocol performance. In this overview we discussed mobility-related aspects of conference signaling and network performance, nomadic user and group management, as well as emerging capabilities of adaptive, scalable video coding. The combination of these ingredients, current or future algorithms and technologies are likely to provide a firm basis for the procurement of widely favored mobile video services.

REFERENCES

Choi, S. J., & Woods, J. W. (1999). Motion-compensated 3-D subband coding of video. *IEEE Transactions on Image Processing*, *8*(2), 155-167.

Compression Links. (2006). Retrieved from http://www.compression-links.info/MPEG-4_AVC_H264

Cycon, H. L., Palkow, M,. Schmidt, T. C., Wählisch, M., & Marpe, D. (2004). A fast wavelet-based video codec and its application in an IP version 6-ready serverless videoconferencing system. *International Journal of Wavelets, Multiresolution and Informational Processing*, *2*(2), 165-171.

Flierl, M. (2003, September). Video coding with lifted wavelet transforms and frame-adaptive motion compensation. *Proceedings of VLBV.*

ITU-T Recommendation. (2000). *H.323: Infrastructure of audio-visual services – Systems and terminal equipment for audio-visual services: Packet-based multimedia communications systems*. Draft Version 4.ITU H.323.

ITU-T Recommendation. (2005). *H.264 & ISO/IEC 14496-10 AVC, Advanced video coding for generic audiovisual services* (version 3)ITU H.264.

Johnson, D. B., Perkins, C., & Arkko, J. (2004). *Mobility support in IPv6*. RFC 3775, IETF.

Koodli, J. (Ed.) (2005). *Fast handovers for mobile IPv6*. RFC 4068, IETF.

Kwon, T. T., Gerla, M., Das, S., & Das, S. (2002, October). Mobility management for VoIP service: Mobile IP vs. SIP. *IEEE Wireless Communications*, pp. 66-75.

Mahy, R., Campbell, B., Sparks, R., Rosenberg, J., Petrie, D., & Johnston, A. (2005). A call control and multi-party usage framework for the session initiation protocol (SIP). IETF Internet Draft - work in progress October 2005.

Marpe, D., Schwarz, D., & Wiegand, T. (2003). Context-adaptive binary arithmetic coding in the H.264/AVC video compression standard. *IEEE Transactions CSVT*, *13*(7), 620-636.

Marpe, D., Wiegand, T., & Sullivan, G. J. (2005, September). The H.264/MPEG4-AVC standard and its fidelity range extensions. *IEEE Communications Magazine.*

Mishra, A., Shin, M., & Arbaugh, M. (2003). An empirical analysis of the IEEE 802.11 MAC layer handoff process. *SIGCOMM Comput. Commun. Rev.*, 33(2), 93-102.

Ohm, J.-R. (2002, July). *Complexity and delay analysis of MCTF interframe wavelet structures*. ISO/IEC JTC1/SC29/WG11, Document M8520.

Ohm, J.-R. (2005, July). *Introduction to SVC extension of advanced video coding*. ISO/IEC JTC1/SC29/WG11, Document N7315. Retrieved from http://www.chiariglione.org/mpeg/technologies/mp04-svc/Ohm

Ostermann, J., Bormans, J., List, P., Marpe, D., Narroschke, M., Pereira, F., Stockhammer, T., & Wedi, T. (2004, April). Video coding with H.264 / AVC: Tools, performance, and complexity. *IEEE Circuits and Systems Magazine, 4*(1), 7-28.

Richardson, I. E. G. (2006). *H.264/MPEG-4 Part 10 Tutorials*. Retrieved from http://www.vcodex.com/h264.html

Reichel, J., Schwarz, H., & Wien, M. (Eds.). (2005, October). *Scalable video coding – Joint draft 4*. Joint Video Team, Doc. JVT-Q201, Nice, France.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., et al. (2002). *SIP: Session initiation protocol*. RFC 3261, IETF.

Rosenberg, J., & Schulzrinne, H. (2002). *Session initiation protocol (SIP): Locating SIP servers*. RFC 3263, IETF.

Schäfer, R., Schwarz, H., Marpe, D., Schierl, T., & Wiegand, T. (2005, July). MCTF and scalability extension of H.264/AVC and its application to video transmission, storage, and surveillance. In *Proceedings of VCIP 2005*, Peking, China.

Schierl, T., Schwarz, H., Marpe, D., & Wiegand, T. (2005). *Wireless broadcasting using the scalability extension of H.264/AVC*. Submitted to ICME 2005, Amsterdam, Netherlands.

Schmidt, T. C., Wählisch, M., Cycon, H. L., & Palkow, M. (2003). Global serverless videoconferencing over IP. *Future Generation Computer Systems*, *19*, 219-277.

Schmidt, T. C., & Wählisch, M. (2003). Roaming real-time applications: Mobility services in IPv6 networks. In *Proceedings of TERENA 2003 Networking Conference*.

Schmidt, T. C., & Wählisch, M. (2005). Predictive versus reactive: Analysis of handover performance and its implications on IPv6 and multicast mobility. *Telecommunication Systems*, *30*(1-3), 123-142.

Schmidt, T. C., Wählisch, M., Cycon, H. L., & Palkow, M. (2006). SIP initiated mobile multimedia group conferencing based on SSM. In *Proceedings of TERENA 2006 Networking Conference*.

Soliman, H., Castelluccia, C., El Malki, K., & Bellier, L. (2005). *Hierarchical mobile IPv6 mobility management* (HMIPv6). RFC 4140, IETF.

Schwarz, H., Hinz, T., Kirchhoffer, H., Marpe, D., & Wiegand, T. (2004, October). *Technical description of the HHI proposal for SVC CE1*. ISO/IEC JTC1/WG11, Doc. M11244. Palma de Mallorca, Spain.

Schwarz, H., Marpe, D., & Wiegand, T. (2004, October). SNR-scalable extension of H.264/AVC. In *Proceedings of IEEE International Conference on Image Processing* (ICIP 2004), Singapore.

Schwarz, H., Marpe, D., & Wiegand, T. (2005, July 6-8). Combined scalability support for the scalable extension of H.264/AVC. In *Proceedings of ICME 2005*. Amsterdam, The Netherlands.

Schwarz, H., Marpe, D., & Wiegand, T. (2006, October 8-11). *Overview of the scalable H.264/MPEG4-AVC extension*. Paper presented at IEEE International Conference on Image Processing (ICIP 2006). Atlanta, GA.

Stockhammer, T., Hannuksela, M. M., & Wiegand, T. (2003). H.264/AVC in wireless environments. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(7), 657-673.

Wedlund, E., & Schulzrinne, H. (1999). Mobility support using SIP. In *Proceedings of the 2nd ACM International Workshop on Wireless Mobile Multimedia* (pp. 76-82). ACM Press.

Wiegand, T., Sullivan, G. J., Bjøntegaard, G., & Luthra, A. (2003, July). Overview of the H.264/AVC video coding standard. *IEEE Transactions CSVT*, *13*(7), 560-576.

Wiegand et al. (2006). *Scalable extension of H.264/AVC*. Retrieved from http://ip.hhi.de/imagecom_G1/savce/

Wikipedia, H.264. (2006). Retrieved from http://en.wikipedia.org/wiki/H264

KEY TERMS

Any Source Multicast (ASM): The distribution of packets from an arbitrary source to a group of receivers. Receivers are addressed by a delocalized group address and remain unidentified by the source.

DCT: Discrete cosine transform.

H.264/AVC: Video compression decompression (codec) international standard.

ITU-T: International Telecommunication Union—Telecommunication Standardization Sector.

ISO/IEC: International Organization for Standardization /International Electrotechnical Commission.

JVT: Joint Video Team.

MPEG: Moving Picture Experts Group.

NAL: Network abstraction layer.

Nomadic User: A term for the user behaviour of switching devices.

Real-Time Transport Protocol (RTP): Internet protocol standard for the transmission of real-time data, providing media sessions, timestamps, order and media-specific metadata encoding.

Session Initiation Protocol (SIP): Internet protocol standard for managing session-based network communication.

Signal to Noise Ratio (SNR): The common quality measure in signal encoding.

Source Specific Multicast (SSM): The distribution of packets from an initially specified source to a group of receivers. Listeners must actively subscribe by address to each source they want to receive traffic from.

SVC: Scalable video codec.

VCEG: Video Coding Experts Group.

Mobile Sports Video with Total Users Control

Dian Tjondronegoro
Queensland University of Technology, Australia

INTRODUCTION

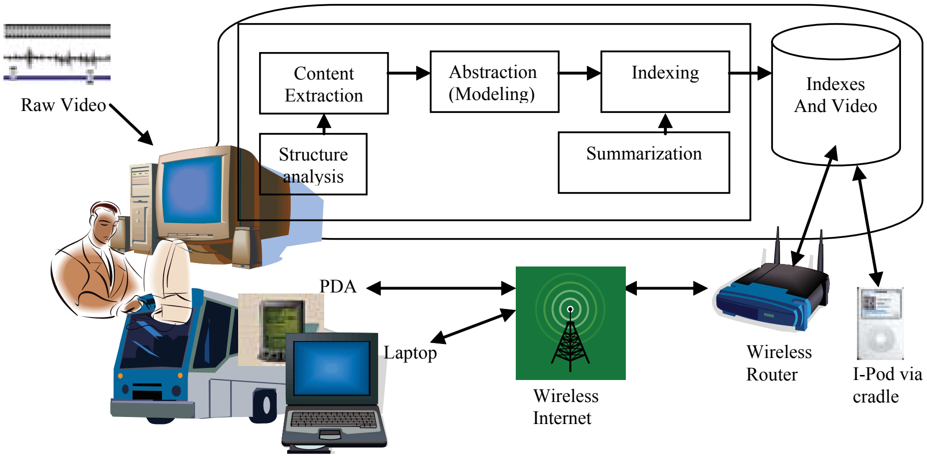
Sports video is very popular thanks to its in-progress (live) information and entertainment values. Many users are motivated to access sports video using mobile devices, since they often cannot watch the game on their sofa due to a busy life and inability to cope with lengthy games. The current generation of mobile video services has only focused on supporting the when and where consumers can watch their favorite sports matches. Since total control over playback and content is neglected, users often have to settle with low-quality videos and static content, which have been pre-processed. This limitation slows down the progress towards an era in which users are comfortable using their mobile devices to enjoy sports broadcasts while gaining total control over what they can watch at their most convenient time and place. In this article, we will describe a mobile video system which offers users full support over the when, where and how they want to watch sports video. The main new features offered are: (1) non-linear navigation within single and/or multiple documents; (2) customizable and personalized summaries; (3) multimodal access and video representation.

BACKGROUND

With the ongoing growth of mobile and hand-held video-enabled devices such as PDAs, smartphones, and iPods, users are increasingly able to afford video on-demand at their most convenience. Sports video is particularly popular for mobile consumption, thanks to its in-progress information (live) and entertainment values. Consumers are at the most convenient level if they have total control over the "when, where and how" they can access the video. The following will describe two scenarios that show how users can benefit from total control over mobile sports video. Figure 1 gives an overview of the descriptions.

- **Scenario 1:** A soccer fan does not want to miss two important live matches, one in the morning (7 a.m.), and another one in the evening (7 p.m.). Thus, he schedules the recording of both broadcasts on his home PC with TV tuner. While the video is being recorded, he wants to watch it on his laptop during a one-hour trip to work. Unfortunately, the bus is full and he has to stand with noisy and squeezed passengers. Thus, he uses his PDA that is connected via wireless network to listen to the audio and occasionally look at the live images. As soon as he arrives at work, he has to attend meetings and finish many tasks, which are due on that day. During this time (8 hours), his PC has finished the match recording and indexing process. At 6:30 p.m. he finishes work, and during his bus-trip from work he must watch the first match while skipping many boring tracks (i.e., those without any interesting events), so that he can watch the first 30 minutes of

Figure 1. Mobile sports video architecture



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the second match live on his laptop (by using wireless broadband to access his PC via remote desktop). As soon as he arrives at home, he must cook dinner, thus he uses his PDA in his pocket to listen to the match. During dinner, he cannot watch the game since it is a special time for his family, hence he settles with audio commentaries being turned on only when there is a goal scored. At 8:30 p.m., he can relax and watch the rest of the live match using the media center in his living room. In this scenario, we assume that all devices are connected with a wireless network and all have access to broadband Internet. This setup allows each device to inform the server (in this case the home PC) the current time-stamp in which the user stops each playback.

- **Scenario 2:** A casual sports viewer does not usually watch live broadcasts since she likes to follow various sports and has a very busy life. However, she does not want to wait and follow static summaries from TV and Internet. Instead, she prefers to be able to watch the personalized summary of each sport video every day or at least three times in a week during her trip from-and-to work. She likes the contents to be pre-recorded on her home PC, indexed, and prepared according to her personal preferences so that she can store them in her iPod. For swimming and other racing sports, she likes to watch every race in real-time while skipping the interviews. For soccer, basketball, and other score-based games, she only wants to see all key events (i.e., goal and goal attempts), as well as any interesting segments in which her favorite players appear. For tennis, badminton and other set-point based games, she wants to see only the key events, such as long rally and service ace, as well as the last portion of play just before each set is won by the player. Due to the unpredictability of public transport (e.g., lighting, crowd, etc.), she needs each segment to be available in multimodal forms so that she can listen or see the key audio or images. Moreover, when she has too many videos to be watched, she needs the video segments to be presented in key frames in order to choose which clips she wants to play.

The two scenarios demonstrate that mobile video interaction increases the necessity for more effective content-based retrieval. Mobile devices have limited capabilities in supporting users watching the full contents of a sports video due to their small screen size and restricted battery life. Since local storage in mobile devices is relatively small, costs of downloading streaming content is also a major issue. Thus, users need to selectively watch particular segments they want to watch to reduce the time and costs of downloading full-video content. Current solutions for streaming video content have not fully exploited the power of content-based indexing to enhance users' experiences. For example, UEFA.com via RealPlayer 10 only allows users to watch a fixed set of soccer video segments which are compiled as a 15 minute highlight. To improve the flexibility of content access, the system must more adaptive to a user's requirements. A study on users' requirements on mobile TV content has found that viewing was most likely to be transient and low commitment, as people are worried about getting too absorbed and are distracted with other tasks (Knoche & McCarthy, 2005). One of the biggest challenges in mobile sports video streaming is how to repurpose TV-quality videos for mobile devices to meet the tolerable downloading time, color depth, and available network bandwidth (Lum & Lau, 2002). Traditional approaches have simply sacrificed the video quality and produce distorted pictures and audio, which are not necessarily acceptable for small screen and speakers. Hence, users should be able to select which key segments (highlights) to watch, thus content-based analysis of sports video is needed to extract the important content automatically. Automatic sports video highlights extraction has been a major issue, which has been addressed by researchers world-wide (Babaguchi, Ohara, & Ogura, 2003; Duan, Xu, Chua, Qi, & Xu, 2003; Ekin & Tekalp, 2003; Rui, Gupta, & Acero, 2000); therefore it will not be the focus of this article.

The scenarios also show the importance of context awareness as an integral feature in mobile computing, which means that applications should react according to the circumstances in which they operate (Wikipedia, 2006). Almost any information available at the time of an interaction can be seen as context information (Korkea-aho, 2000), thus its long list can be categorized into: user-, physical-, computing- and time- contexts (Chen & Kotz, 2000).

Based on these discussions, Table 1 summarizes the requirements for sports video viewing on mobile devices. In this article, we will propose a content-based video indexing and retrieval that supports a total control over the video accessing (i.e., M1-M4) and how it can meet all the other requirements (i.e., T1-T4, and S1-S4). To demonstrate the look-and-feel of the system, we have implemented a Web-based video retrieval system that can be accessed by desktop and mobile devices; thus supporting multi-platform access (S1). Figure 2 depicts the system's interface.

TOTAL CONTROL ON PLAY-BACK

It should be noted that all video playback tools, such as Windows Media Player and Apple QuickTime, support total control over: *play*, *pause*, and *rewind* during live streaming video (T1). Fast forward is only available when the future content is already available; for example, play-back is delayed after the video is recorded (T2). While play and pause are very straight forward operations, fast forward and rewind should be achieved in a more elegant manner, rather than

Figure 2. Web-based video retrieval system accessible for desktop and mobile devices

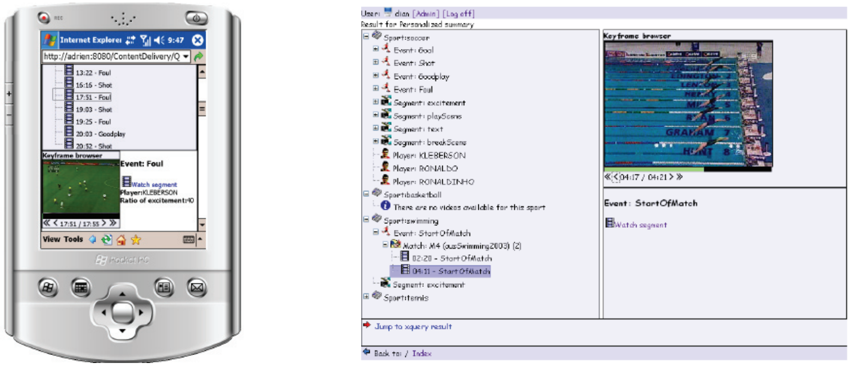

Internet Explorer 9:47
http://adrien:8080/ContentDelivery/Q
13:22 - Foul
16:16 - Shot
17:51 - Foul
19:03 - Shot
19:25 - Foul
20:03 - Goodplay
20:52 - Shot
Keyframe browser
Event: Foul
Watch segment
Player:KLEBERSON
Ratio of excitement:40
« < 17:51 / 17:55 > »
View Tools
Pocket PC
User: dian [Admin] [Log off]
Result for Personalized summary
Sport:soccer
Event: Goal
Event: Shot
Event: Goodplay
Event: Foul
Segment: excitement
Segment: playScene
Segment: text
Segment: breakScene
Player: KLEBERSON
Player: RONALDO
Player: RONALDINHO
Sport:basketball
There are no videos available for this sport
Sport:swimming
Event: StartOfMatch
Match: M4 (ausSwimming2003) (2)
02:20 - StartOfMatch
04:11 - StartOfMatch
Segment: excitement
Sport:tennis
Keyframe browser
« < 04:17 / 04:21 > »
Event: StartOfMatch
Watch segment
Jump to xquery result
Back to: / Index

Table 1. Requirements for most-convenient mobile video viewing

WHEN (**T**emporal)	WHERE (**S**ituation)	HOW (Access **M**ethod)
T1: During (real-time): contents are presented as-it-is. Indexing can benefit from Web sites that offer real-time content annotation	**S1:** Seamless multi-platform access: users share familiar GUI and data on desktop as in mobile devices	**M1:** Total control on the video playback: pause, rewind, fast-forward, and so forth
T2: After (Recorded): there is time for processing content extraction and indexing. Often we don't want a summary of the content (ruining the excitement).	**S2:** *User context*: profile (identity), activity, and varying pace of interaction, Surroundings (including people nearby), social situations, schedules and agendas.	**M2:** *Non-linear (hyper) navigation within single and/or multiple documents
T3: Before (Preview or build-up): Team and player introductions, score predictions, supporters' comments	**S3:** *Physical context*: lighting, noise, temperature	**M3:** *Customizable and personalized summaries: users should not write their own queries.
T4: *Time context*: time of a day (morning/night), week, month, season, and so forth	**S4:** *Computing context*: adaptive to hardware (platform) and network resources	**M4:** *Multimodal access: key representations for low-bandwidth, using sound/image

** M1 – M4 represents requirements for total control over the playable contents*

just jumping several frames. In fact, content-based intelligent fast forward should support smart skimming, which allows users to skip uninteresting segments without missing too much of information. This means that a user would be able to recognise the aim of the video's contents via a shorter version of the original video. Intelligent fast-forward (or skimming) supports a varying pace of interaction, which is often determined by users' contexts, including activity, schedules and agenda (S2). During mobile video consumption, users will benefit from fast skimming especially if the current content is boring or when the playing device is running out of battery, which part of computing context (S4).

In order to achieve a successful skimming, the system should exploit the fact that users unintentionally embed their understanding and interests of the video content through their interaction with computers. Thus, video skimming for future users can be based on modeling previous users' experiences and attention with the video. One method is to use *browsing behavior* since intelligent skimming can be achieved by delivering shorter video clips which are personalized based on users' preferences. Common tracking of video usage can be achieved by examining a usage log to determine the number of times a video (segment) has been accessed. For example, a shot rank calculation framework can be used to measure the subjective importance of video shots by unifying low-level video analysis and user-browsing-log mining (Yu, Ma, Nahrstedt, & Zhang, 2003). Since usage log is not always sufficient to reveal user interests, a statistical model such as Hidden Markov Model can be used to analyze the potential browsing states while users are watching videos, such as "aimless browse," "looking for something," and "found what I wanted" (Syeda-Mahmood &

Ponceleon, 2001). Thus, interesting segments can be defined as the segments "played" by users in an "interested" state. Unstructured video contents should be therefore managed by standard meta-data. Another method is to use *user attention model* to presents contents that would be generally interesting for users. While watching a video sequence, human attention is generally attracted by various information channels including image sequence, audio track, and text information. A complete user attention model should be fused using machine learning or linear combination from the user attention model of each modality. For example, by combining audio and visual models, video skims can be produced without interrupting sentences in the audio channel (Ma, Lu, Hong-Jiang, & Li, 2002) developed. Thus, video skims can generally generated by selecting particular visual objects and audio keywords (Smith & Kanade, 1998). For example, as soccer video contains many global shots, users will find it easier to follow the action if the regions of interest are zoomed-in (by cropping) on small devices in real time (Seo & Kim, 2006).

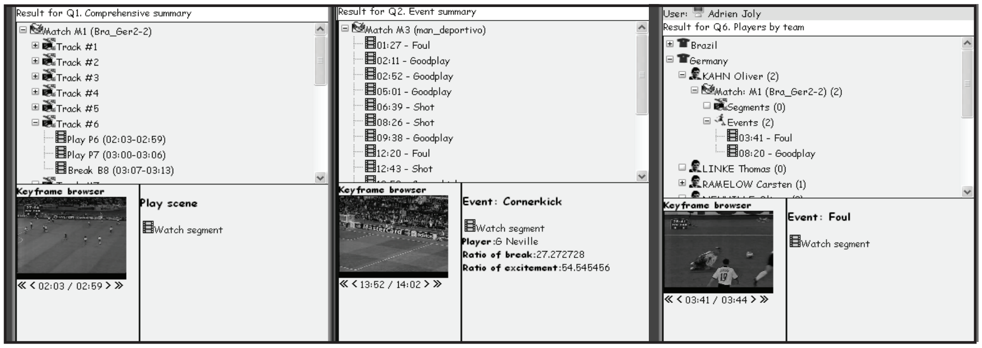
NON-LINEAR NAVIGATION

Current mobile video players have yet to fully support non-linear navigation of video contents within a single and/or multiple documents. This feature is particularly useful if users are presented with a large collection of videos. For example, users can benefit browsing on the contents within a single video according to the topics (e.g., soccer goals), or within multiple documents according to subjects, topics classification and taxonomy. To support this feature, content extraction and summarization processing are required. In this section, we will describe some video summarization schemes that enable non-linear navigation.

- **Comprehensive Summary:** We have applied a hierarchical structure to organize a sports video summary that consists of integrated plays, breaks and highlights as shown in *Figure 3a*. Using this summary structure, users can easily choose to browse a sports video either by play-break sequences (like audio tracks), or collection of highlights (based on the categories such as goals and fouls). When a particular collection is selected, users can select the particular highlight segment. Each highlight segment will consist of *play* and *break* shots. Most sport viewers prefer to focus their attention to events within play segments. It is due to the fact that most sport videos contain many events that cause a game to stop, such as foul, goal celebration and end of playing period in soccer. In most cases, even a sports fan does not want to spend their time waiting for the game being resumed again. On the other hand, if users prefer to browse by sequences, they can check whether the sequence contains a highlight. Thus highlight can be a subset or the whole length of a sequence. Users can watch the entire sequence or watch the highlight only for a shorter version. Based on a user study reported in Tjondronegoro, Chen, & Pham (2004), we found that most sports fans would rather watch the entire sequence to fully understand the context of the highlight, while most casual viewers chose to watch just the highlights to save some viewing time. However, both user classes liked the idea of skipping play-break sequence, just like they can skip a track in audio compact disc (CD), since each sequence is "self-consumable." Generally, a play-break sequence contains events which can be classified based on the sports genre, while play segments describe the cause (of each event), and break segment describes the outcome (of each event).

  This model has some obvious benefits. First, users can watch all play and break scenes or just the ones that

Figure 3. (a) comprehensive summary, events summary and team summary; (b) user preference-based summary; (c) player's summary; (d) favorite match summary



Result for Q1. Comprehensive summary
⊟ Match M1 (Bra_Ger2-2)
  ⊞ Track #1
  ⊞ Track #2
  ⊞ Track #3
  ⊞ Track #4
  ⊞ Track #5
  ⊟ Track #6
    Play P6 (02:03-02:59)
    Play P7 (03:00-03:06)
    Break B8 (03:07-03:13)

Keyframe browser
« < 02:03 / 02:59 > »

Play scene
Watch segment

Result for Q2. Event summary
⊟ Match M3 (man_deportivo)
  01:27 - Foul
  02:11 - Goodplay
  02:52 - Goodplay
  05:01 - Goodplay
  06:39 - Shot
  08:26 - Shot
  09:38 - Goodplay
  12:20 - Foul
  12:43 - Shot

Keyframe browser
« < 13:52 / 14:02 > »

Event: Cornerkick
Watch segment
Player:G Neville
Ratio of break:27.272728
Ratio of excitement:54.545456

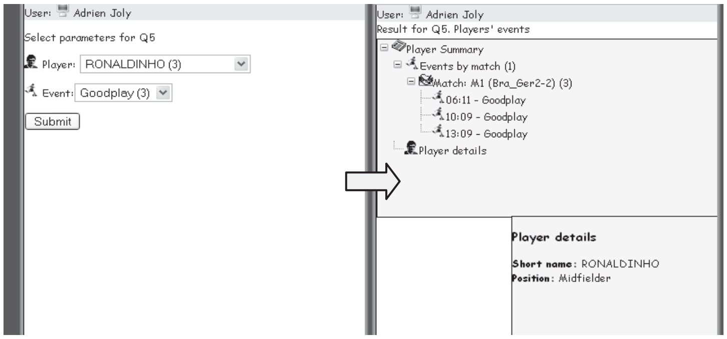
User: Adrien Joly
Result for Q6. Players by team
⊞ Brazil
⊟ Germany
  ⊟ KAHN Oliver (2)
    ⊟ Match: M1 (Bra_Ger2-2) (2)
      ☐ Segments (0)
      ⊟ Events (2)
        03:41 - Foul
        08:20 - Goodplay
  ☐ LINKE Thomas (0)
  ⊞ RAMELOW Carsten (1)

Keyframe browser
« < 03:41 / 03:44 > »

Event: Foul
Watch segment

(a)

Figure 3. continued



User: Adrien Joly

Select parameters for Q5

Player: RONALDINHO (3)

Event: Goodplay (3)

User: Adrien Joly

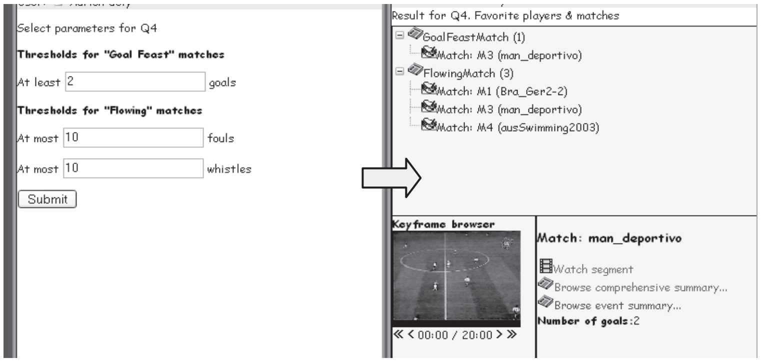
Result for Q5. Players' events

- Player Summary
  - Events by match (1)
    - Match: M1 (Bra_Ger2-2) (3)
      - 06:11 - Goodplay
      - 10:09 - Goodplay
      - 13:09 - Goodplay
  - Player details

Player details

Short name: RONALDINHO
Position: Midfielder

(b)



Select parameters for Q4

Thresholds for "Goal Feast" matches

At least 2 goals

Thresholds for "Flowing" matches

At most 10 fouls

At most 10 whistles

Result for Q4. Favorite players & matches

- GoalFeastMatch (1)
  - Match: M3 (man_deportivo)
- FlowingMatch (3)
  - Match: M1 (Bra_Ger2-2)
  - Match: M3 (man_deportivo)
  - Match: M4 (ausSwimming2003)

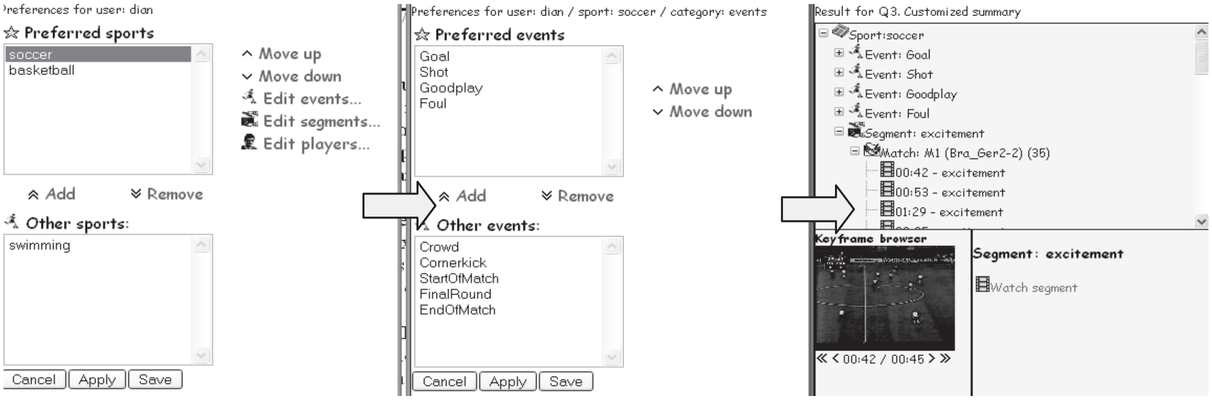
Keyframe browser

Match: man_deportivo

Number of goals: 2

Submit
Watch segment
Browse comprehensive summary...
Browse event summary...
00:00 / 20:00

(c)



Preferences for user: dian

Preferred sports

soccer
basketball

Other sports:

swimming

Preferences for user: dian / sport: soccer / category: events

Preferred events

Goal
Shot
Goodplay
Foul

Other events:

Crowd
Cornerkick
StartOfMatch
FinalRound
EndOfMatch

Result for Q3. Customized summary

- Sport:soccer
  - Event: Goal
  - Event: Shot
  - Event: Goodplay
  - Event: Foul
  - Segment: excitement
    - Match: M1 (Bra_Ger2-2) (35)
      - 00:42 - excitement
      - 00:53 - excitement
      - 01:29 - excitement

Keyframe browser

Segment: excitement

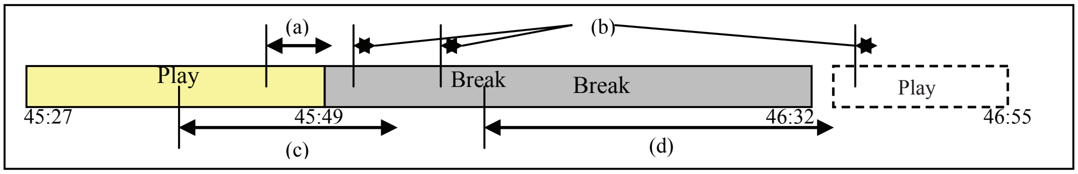
Move up
Move down
Edit events...
Edit segments...
Edit players...
Add
Remove
Cancel
Apply
Save
Move up
Move down
Add
Remove
Cancel
Apply
Save
Watch segment
00:42 / 00:45

(d)

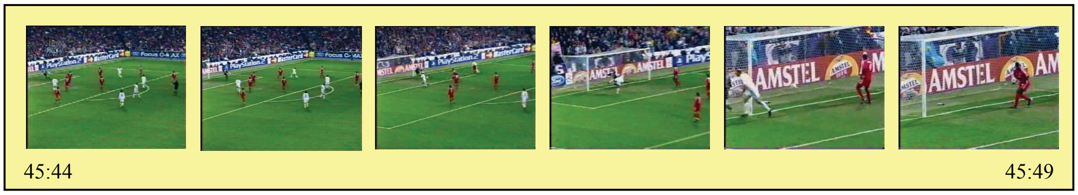
have a certain number of highlights. Second, users can refer back to the whole play or break scene and thus answer: What happens before and after a highlight? Or what causes a highlight? Third, the model lets viewers have one or more highlight collections for a sports video and structures them in a hierarchical scheme. Thus, users can build their own highlight collection on top of existing or system-generated collections.

- **Events Summary:** Event-based indexing can be used as the potentially most suitable indexing technique for sport videos, since sport highlights on TV, magazine or Internet are commonly described using a set of events, particularly the important or exciting events (i.e., key events or highlights). The first reason is the fact that a sports match can be naturally decomposed into specific events. For example, soccer videos may

Figure 4. Multimodal representation of sports video



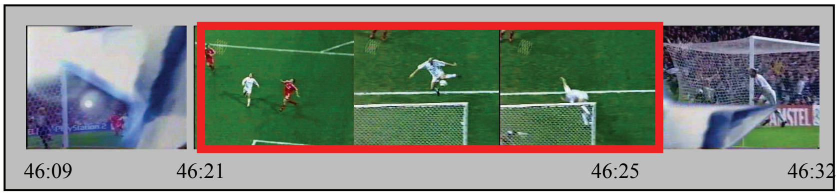
P-B Sequence Containing Goal



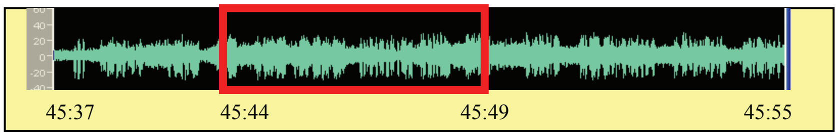
(a) Last near-goal segment

45:52
45:58
46:38

(b) Face and inserted texts



(c) Slow-motion replay scene (highlighted-portion is associated with excitement)



(d) Ecitement during play (highlighted-portion is associated with the last near-goal before goal)

contain players kicking the ball, scoring goals, and so on. Secondly, sport viewers remember and recall a sport match based on the events, especially the most exciting events. For instance, viewers will remember that a goal is scored during the first playing period of a soccer match after watching the video. Thus, when he/she wants to play that goal event again, the query can be based on the particular period of play within the soccer match. Thirdly, events can assemble, describe and summarize specific audiovisual features in soccer videos. Thus, events can serve as an effective bridge between low-level features and high-level features in sports videos. Finally, events occurrence and its order during a specific soccer/sport match can be predicted based on the specific domain knowledge and therefore can be detected automatically.

- **Players by Team Summary:** Most sport fans support particular teams and players. Thus, this summary presents all video segments in which players that belong to particular teams appear. This summary has an apparent entertainment value since users can easily see all the videos in which their favorite teams are involved. Figure 3(a) depicts the screen captures of players by team summary and match events summary.

Players by team summary can be used to construct a *preview* clip (T3) that shows the performance of two competing teams. It is obvious that non-linear navigation of video content supports a varying pace of interaction (S2). Users do not need to follow the standard pace of each match since they can easily browse certain topics or events, move between contents easily within one video and between videos interchangeably. Moreover, depending on users' regularity in accessing the summarized contents (T4), they can choose to watch summaries at any particular time. For example, after a competition like FIFA (soccer) World Cup, users can view all events that belong to the team and players that have competed.

CUSTOMIZED PERSONAL SUMMARIES

While using mobile devices, users have a limited capability in entering complex queries. Thus, content-based and personalized summarization that creates browseable, dynamic event structures will make users feel in total control. Using customized and parameterized summaries, we can support users' context including personal profile (S2). Personalized summary can be constructed using:

- **Parameterized Queries:** Users need to understand the query's syntax language and the data indexing schema before they can write queries of their retrieval requirements. A better alternative is to assist users in selecting what they want to watch by constructing "*queries-generated*" dynamic summaries, event structures, and customized summaries. Thus, users only need to "fill-in-the-gaps."
- **Users' Preferences:** Users should be able to explicitly specify what they want to store and watch.
- An automatic personal summary can be constructed based on search (retrieval) and accessing history (log). This has been discussed previously in the "total playback control" section.
- **Annotation and Metadata:** Users can share their preferences for future searches. For example, each individual user may have favorite and past searched and played clips. Based on annotation by community, popular and top-rated clips can be determined.
- **Parameterized Summary 1—Player's Summary:** (Figure 3b)This query is able to list a certain player's (e.g., Ronaldo) details, such as full name, short name, position, and any event segments that related to him regardless of in which match. It determines all specified event segments in all matches in which the selected player appeared and displays a count on that event (e.g., goals) too. This query facilitates a user tracking a certain player's performance, such as how many goals he scores and how many fouls he made, in all matches in the video database.
- **Parameterized Query 2—Favorite Players and Matches Summary:** (Figure 3c) In this query, users can specify the type of their favorite match and/or players. For example, some users only like to watch a "goal feast" match, which is a match with > N number of goals.
- **User Preference-Based Summary:** (Figure 3d) Users can explicitly choose their favorite type of events, players and segment in this preference. This summarization is particularly useful when users want to specify the total duration of the summary while having total control on the segments, events and players that they will be able to watch.

MULTIMODAL ACCESS

Since video contains a synchronized content of audiovisual modalities, it is naturally larger than other media types such as image and sound. To reduce bandwidth and produce a more compact representation, video can be represented by "lighter" media and smaller clips, including scenes, shots, frames, images (mosaic, multi-frames capture), sound, and text. Multimodal access supports physical context (S3); for example, users may prefer to watch (image-only) for a noisy environment (unless users have headphones), or "audio-only" while standing on a crowded public transportation. Video playback can also be difficult to watch when the lighting is too bright. Multimodal also allows users to multi-task in a busy lifestyle (S2), as described in Scenario 1. For instance, users may only be able to listen to the audio while looking at the full video during key events only. Multimodal video indexing will be adaptive to different hardware and network resources. Unlike most of the current solutions for mobile video playback, which sacrifice video quality over lower-bandwidth and smaller screen, multimodal access can allow users to preview a particular video segment (in the form of image and text) and choose to play a particular clip in a high-quality format.

Our system constructs the sports video index using two main abstraction classes, namely, *segment* and *event*. Each segment is instantiated with a unique key of segment ID into either video, visual, or audio segment. An event can be instantiated into generic (e.g., interesting event), domain-specific (e.g., soccer goal), or further-tactical (e.g., soccer free kick) semantics. Events and segments are chosen as they can provide an effective description for many sport games. For example, most users will benefit from watching soccer goals as the most celebrated and exciting event. Segments can be used as the text-alternative annotations to describe the goal. As shown in Figure 4, the last near-goal segment in a play-break sequence containing goal describe *how* the goal was scored. Face and text displays can inform *who* scored the goal (i.e., the actor of the event) and the updated score. Replay scene shows the goal from different angles to further emphasize the details of how the goal is scored. In most cases, when the replay scene is associated with excitement, the content is more important. Excitement during the last play shot in a goal is usually associated with descriptive narration about the goal. In fact, we (humans) often can hear a goal without actually seeing it.

CONCLUSION

In this article, we discussed some methods that enable users to have total control over the contents of mobile sports video access. The key requirement is content extraction and indexing, as the system can present results according to how users want it, anywhere and at anytime. Full play-back control is achieved by supporting the traditional play and pause operations, plus an intelligent fast-forward or skimming that helps viewers to get an overview of the video's contents. Non-linear navigation within single and/or multiple documents is supported by comprehensive summary of a single match containing play-break and events segments summary in a game containing the highlights, as well as a summary that puts together all segments from the database in which a users' favorite players and teams appear. Customizable and personalized summaries can be achieved manually by allowing users to enter parameters on dynamic summaries and specify their preferences in a stored user profile, or automatically by allowing the system to learn from the usage history and behavior. Multimodal access is achieved by representing video with other modalities such as key frame or image and audio.

ACKNOWLEDGMENT

The authors would like to sincerely thank Adrien Joly who developed the Web-based system during his stay at QUT for exchange program with INSA de Lyon, France.

REFERENCES

Babaguchi, N., Ohara, K., & Ogura, T. (2003). *Effect of personalization on retrieval and summarization of sports video*. Paper presented at the Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia.

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research*. Dartmouth College.

Duan, L.-Y., Xu, M., Chua, T.-S., Qi, T., & Xu, C.-S. (2003). *A mid-level representation framework for semantic sports video analysis*. Paper presented at the ACM MM2004, Berkeley, USA.

Ekin, A., & Tekalp, M. (2003). Automatic soccer video analysis and summarization. *IEEE Transaction on Image Processing*, *12*(7), 796-807.

Knoche, H., & McCarthy, J. D. (2005). *Design requirements for mobile TV*. Paper Presented at the 7th International Conference on Human Computer Interaction with Mobile Devices & Services, Salzburg, Austria.

Korkea-aho, M. (2000, April 25). *Context-aware applications survey*. Retrieved July 2006, from http://users.tkk.fi/~mkorkeaa/doc/context-aware.html

Lum, W. Y., & Lau, F. C. M. (2002). A context-aware decision engine for content adaptation. *Pervasive Computing*, *1*(3), 41-49.

Ma, Y.-F., Lu, L., Hong-Jiang, Z., & Li, M. (2002). *A user attention model for video summarization*. Paper presented at the 10th ACM International Conference on Multimedia, France.

Rui, Y., Gupta, A., & Acero, A. (2000). *Automatically extracting highlights for TV baseball programs*. Paper presented at the ACM International Conference on Multimedia, Marina del Rey, California.

Seo, K., & Kim, C. (2006). *A context-aware video display scheme for mobile devices*. Paper presented at the Multimedia on Mobile Devices II, San Jose, CA.

Smith, M. A., & Kanade, T. (1998). *Video skimming and characterization through the combination of image and language understanding*. Paper presented at the Content-Based Access of Image and Video Database.

Syeda-Mahmood, T., & Ponceleon, D. (2001). *Learning video browsing behavior and its application in the generation of video previews*. Paper presented at the Ninth ACM International Conference on Multimedia, Ottawa, Canada.

Tjondronegoro, D., Chen, Y.-P. P., & Pham, B. (2004, October-December). Integrating highlights to play-break sequences for more complete sport video summarization. *IEEE Multimedia*, 22-37.

Wikipedia. (2006, June 7). *Context awareness.* Retrieved July 2006, from http://en.wikipedia.org/wiki/Context_awareness

Yu, B., Ma, W.-Y., Nahrstedt, K., & Zhang, H.-J. (2003). Video summarization based on user log enhanced link analysis. Paper presented at the 11th ACM International Conference on Multimedia, Berkeley, CA.

KEY TERMS

Break: A break is similar to play, except it happens when the game is stopped or paused due to specific reasons such as foul and goal.

Browsing Behaviour: Browsing behaviour is the way users typically look for the desired information.

Event: An event is a special type of video segment that contains a particular theme or topic such as a soccer match, goal, and foul. An event includes (a) temporal textures such as flowing water: indefinite spatial and temporal type, (b) activities such as a person walking: temporally periodic but spatially restricted and (c) isolated motion events such as smiling: no repeat either in space or in time.

Multimodal: Multimodal means that a system utilizes more than one mode or modality (as of stimulation or treatment) to transfer information. This term is often called multimedia since video appeals to more human senses than an image.

Object: An object is either an abstract or visible entity that appears in a video segment; therefore it can be regarded as the actor of the segments and events.

Play: A play is a specific type of video segment in a sports match, which shows events when the game is still flowing, such as when the ball is being played in soccer.

Preview Clip: A preview clip is a prepared clip that shows the contents of a future program by collating existing clips which have similar topics of interests. For example, a soccer match preview can show preview clips that contain the key players from each team.

Queries Generated Summary: This is the type of summary that is not predefined statically or prepared manually by collating certain video segments. Instead, it is constructed using a query formula on video indexes.

Segment: A segment is an audio, visual, or audiovisual portion of video that contains index-able contents such as whistle, slow motion replay, close-up players' face and near goal area.

User Attention Model: User attention model is the predictable area of interests from users while watching each video modality, such as a particular spatial area in an image.

Mobile Telephony in Sub-Saharan Africa

Princely Ifinedo
University of Jyvaskyla, Finland

INTRODUCTION

A mobile telephone is a telecommunications device that connects its user to a network using a wireless radio wave transmission technology. In some parts of the world, mobile phones are known as cellular phones. Mobile telephones were first introduced in the mid-1980s (Marcussen, 2002; Sadeh & Sadeh, 2002; Sarker & Wells, 2003). Mobile telephony is diffusing globally due to a variety of reasons, including cost advantages in setting up the system compared to landlines, its small-sized nature, portability, and its ability to foster and enhance social relationships, among others (Plant, n.d.; Marcussen, 2002; Sadeh & Sadeh, 2002; Sarker & Wells, 2003; ITU, 2004; Anonymous, 2006). According to reports by ITU (2004), the percentage of total telephone subscribers that are mobile telephone subscribers has been increasing over the last five years. In 2005, mobile telephone subscribers were approximately 62% of total telephone subscribers for the five regions of the world (see Table 1).

The data in Table 1 shows that Africa has the lowest connectivity rate (i.e., 9 per 100 people), but with the highest number of mobile telephone users as percentage of the total telephone subscribers for 2004 (WEF, 2003; ITU, 2004). The current trend suggests that this will be the case for some time (BBC News, 2002a; WEF, 2003; ITU, 2004). For example, BBC News (2002a) reported that "the popularity of wireless communication is soaring. More mobiles were connected in the past five years than landlines installed in the last century." Similarly, a recent report by the World Economic Forum (WEF, 2003) on ICT status in Africa indicated that the number of mobile telephone subscribers in Africa now numbers around 23 million and the demand is expected to grow in the future. Furthermore, the WEF report also provided the growth estimate for other information and communication technology (ICT) products in Africa which is reproduced in Figure 1. Clearly, the number of mobile telephone users is the highest of the three ICT products compared.

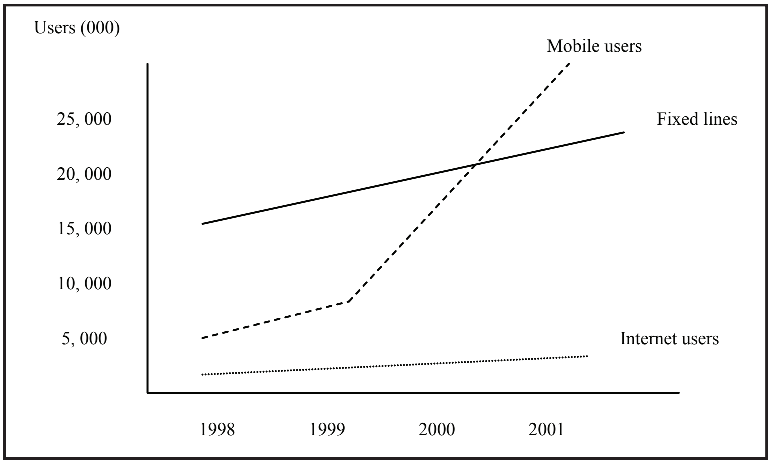
Given the encouraging statistics seen with mobile telephones user in Africa, the following questions are posed:

Table 1. Mobile telephony (cellular) subscribers per 100 people 2004 (ITU, 2004)

Region	CAGR	Per 100 2004	As % of total telephone subscribers 2004
Africa	59.7	9.14	75.1
Americas	22.9	42.74	55.8
Asia	34.4	18.72	56.8
Europe	25.9	71.61	63.6
Oceania	20.7	61.71	59.2

** CAGR (Compound Annual Growth Rate)*

Figure 1. Growth in the number of ICT users in Africa



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- Can Sub-Saharan African countries exploit the diffusion of mobile telephones to create or engage in m-business?
- Why would such a facility be useful for them?

THE REGION OF SUB-SAHARAN AFRICA

Africa, with its population of about 880 million people, is the poorest continent in the world (World Bank, 2001a, 2001b; Ifinedo, 2005a, 2005b; CIA, 2005). In terms of geography, Africa tends to be described as comprising two regions—North Africa and SSA. The northern part is comparable to the Middle East economically and culturally (Ifinedo, 2005b). Further, South Africa (also known as the Republic of South Africa [RSA]) tends to be excluded from the rest of SSA because of its relative high socio-economic indicators. In brief, SSA is associated with poverty, a high illiteracy rate, civil strife, and chronic under-development (World Bank, 2001b; ITU, 2004; CIA, 2005). In this article, the northern part of Africa and RSA are excluded, because these parts of Africa have better indicators for ICT use and are richer than the rest of the continent (Ifinedo, 2005b). Put differently, the conditions in SSA are different from those in the excluded regions, and the region of SSA typifies perceptions of Africa more than do the excluded regions. For example, in a recent report, the World Bank (2001b, p. 38) states: "In Africa, slow growth increased both the share and the number of the poor over the 1990s; Africa is now the region with the largest share of people living on less than US$1 per day." The conditions in SSA, it appears, are more consistent with the foregoing observations than for any of the countries in North Africa and RSA.

With about 10% of the world's population living in the SSA (CIA, 2005), it is sad to note that only 0.2% of the world's one billion fixed telephone lines are located in the region (WEF, 2003, p. 2). Indeed, the current statistics for Africa with regard to the pervasion of mobile telephony in the region is welcoming. However, one of the main questions for policymakers in the region is: how should the advent and spread of mobile telephony in the region be exploited to improve the livelihoods of its inhabitants? Answers to such a question are pertinent because development research and reports have suggested that ICT adoption, diffusion, and usage in the developing countries (including SSA) could hasten socio-economic development in such disadvantaged regions of the world if concerted efforts are made to harness the power of such ICT products (Avgerou, 1998; Castells, 1999; Singh, 2000; G8 DOT Force, 2001; UN ICT TASK Force, 2004; Ifinedo, 2005b). Recently, during an ICT task force meeting at UN headquarters in New York, the UN Secretary General Kofi Annan remarked: "Information and communication technologies can help us turn this potential [of using ICT for socio-economic development] into concrete opportunities that will help the poor work their way out of poverty" (BBC News, 2002a). The UN Secretary General added: "It is not, of course, a magic formula that is going to solve all the problems, but it is a powerful tool for economic growth and poverty eradication, which can facilitate the integration of African countries into the global market."

Along a similar line of reasoning, Singh (2000) discussed how the use of ICT products in developing nations can improve the chronic socio-economic conditions in such regions. He stated, "For firms [and people] in the developing countries dealing in the global market some of the common problems are [as follows:] lack of market knowledge, poor communication, cumbersome procedures, delays and uncertainties in supply, poor quality and excessive stock." He noted that e-commerce (the buying and selling over the Internet) "can help solve some of these through better knowledge management, communication and automated supply procedures leading to higher profits and enhanced competitiveness" (p. 23). It is important to add that the foregoing issues and problems identified by Singh are inhibitors to socio-economic development in developing parts of the world even when the emphasis or bias is not the global market. In other words, poor supply chain management and poor organization could stifle economic development in local trade as well when information does not get to the people that require it.

At this juncture, it is vitally important to investigate if the ongoing positive pervasion of mobile telephony in SSA can facilitate m-business among the people and businesses in the region. Next, we discuss the concepts of mobile business at a general level, and present a summary of how such concepts have been implemented in one SSA country—Senegal—to improve the lot of some citizens in that country.

WHAT IS M-BUSINESS?

Simply, m-business, which is also know as mobile business, can be defined as doing business using wireless services. This is the mobile equivalent of e-business. A clear picture of m-business emerges by understanding e-business. Zwass (1996) defines e-business as "the sharing of business information, maintaining business relationships, and conducting business transactions by means of the telecommunication networks." Similarly, m-business can be defined as the sharing of business information, maintaining business relationships, and conducting business transactions using a wireless radio wave transmission technology. Just as researchers (e.g., Turban, Lee, King, & Chung, 2000) have argued that there is a difference between e-commerce (buying and selling of good and services online) and e-business (a broader term that includes e-commerce and the servicing of customers, collaborating with entities both within an organization and outside it), this article distinguishes between m-business and m-commerce. Broadly speaking, m-commerce is the next-generation e-commerce; it refers to the buying and selling of goods and services through a network of wireless radio wave transmission technology (Sadeh & Sadeh, 2002). The buyers and sellers may use mobile phones (cellular telephones) and/or personal digital assistants (PDAs). Wireless application protocol (WAP) is the emerging technology behind m-commerce and m-business (Anonymous, 2006).

Further, many handset manufacturers, including Nokia, Motorola, and Ericsson, having realized the potential for m-business and m-commerce worldwide, have initiated partnerships with carriers such as AT&T Wireless to develop WAP-enabled smart phones. Another telecommunications industry specification that is increasingly being mentioned in connection with the emerging m-business initiatives is Bluetooth. Essentially, this technology describes how mobile or cellular phones, PDAs, and computers can easily be interconnected using a short-range wireless connection (Anonymous, 2006). M-business and m-commerce issues seem to be diffusing at different rates across the world. Europe and some parts of Asia (notably, South Korea and Japan) appear to be leading the rest of the world on such fronts (Marcussen, 2002; Sadeh & Sadeh, 2002; McKinsey Quarterly, 2006). As would be expected, the diffusion of m-business in SSA is marginally low. However, we present a case study of m-business in Senegal in the next section.

M-BUSINESS IN SSA: A CASE STUDY

M-business in Africa is beginning to be discussed, and development from Senegal seems to be pioneering such initiatives in the SSA region (e.g., Annerose, 2001). Mobile services have been used to improve commercial activities in the country (Annerose, 2001; Manobi, 2003). Like many countries in SSA, Senegal is poor: about 70% of its population lives in rural communities where most of them make their living by farming and fishing (Annerose, 2001; CIA, 2005). However, fallouts from the increasingly globalized world, Annerose (2001) noted, have reached the peoples of SSA, including Senegal. He comments:

> *Like many developing countries, Senegal in recent decades has deregulated and liberalized its economy to adapt to the demands of globalization. In rural areas, this has brought the gradual disappearance of government mechanisms that managed production and prices. Small-scale farmers and fishermen have had great difficulty adapting to a purely market-driven economy, and rural incomes have dropped. One of the key challenges is that farmers and fishermen have no way of determining market prices before they sell their crop or catch to middlemen, many of whom take advantage of this ignorance and offer prices much lower than market prices.*

To assist, the fishermen and farmers in one Senegalese town, Niayes, sell their wares at lower losses, compared to previous times when huge losses were incurred due to the poorly organized supply chains that are commonly seen with businesses in developing parts of the world such as SSA; Manobi, a local private telecommunications company, came up with an initiative using mobile technology to help them gather information about market prices. The first phase of the project commenced in 2001; it involved Manobi and some fruit and vegetable farmers of Niayes. The chief executive officer of Manobi, Annerose, reported that as a result of using the mobile phone-based market price system, farmers were able to increase their prices by over 50% (Annerose, 2001). In summary, farmers use their mobile phones from their homes and farms to check price levels, and the information they garner permits them to make better decisions regarding what to sell, when to sell, and even what to grow. BBC News (2002b) reports: “Even though Manobi is only being tested [in 2001], it is already having an effect on the way farmers grow crops.” The report continued by citing quotes from the CEO of Manobi who asserted that: “For a farmer it is very interesting to note that price is not something stable; a price is living data and is changing very quickly.” (Figure 2 shows a farmer using a mobile phone from his farm.)

Following the success of the pilot project, the initiative expanded, and in 2003, Manobi (the firm)—in partnership with three local fishing unions, two telecommunications companies (Alcatel and Sonatel), and the Canadian International Development Research Center (IDRC)—started another project that aims to support the livelihoods and improve the safety of Senegalese fishermen through the provision of timely information on market prices, weather reports, and other information services via mobile phones using WAP and SMS (short messaging services) technologies. The

Figure 2. An African vegetable farmer using a mobile phone (Photo Source: Messrs Ololube, & Ifinedo)

fishermen are provided with training on how to retrieve the information they require (Manobi, 2003). In brief, the project uses three data collectors who are stationed in markets of the participating cities and towns, including Dakar and Kayar. The information they collect is captured in a Psion computer and is then transmitted by a mobile phone to a central database and Web site. This arrangement ensures that market information is available real time to fishermen. In addition, the fishermen also get access to up-to-date weather reports. The project has been a success, and several individual users (fishermen and artisans) joined by the end of 2003 with a higher number of subscribers expected in the future. BBC News (2002b) also reported:

> *Now Manobi is talking to professional organisations that represent more than 250,000 people who work in Senegal's agricultural industry. Prices [of using the services] are kept low and farmers pay for the service as part of a deal between Manobi and the national telephone company.*

The future of the initiative seems to be secured because Manobi, the national government of Senegal, and other private-sector partners are developing a detailed plan to extend the service to farmers and fishermen across Senegal (Annerose, 2001; Manobi, 2003).

THE USEFULNESS AND CHALLENGES OF M-COMMERCE DIFFUSION IN SSA

To some extent, some aspects of the sorts of problems hampering commercial activities in developing countries that Singh (2000) discussed in his report can be ameliorated through the engagement of m-business. For example, m-business could provide a good opportunity to solve entrenched problems of poor supply chain management that manifest in several ways, including a lack of market knowledge, uncertainties, and/or instability in supply due to poor communication systems and excessive stock. When these kinds of problems are tackled, it is likely that farmers, fishermen, and other artisans that require access to better information can improve their competitiveness, which will invariably lead to higher profits for them and subsequently a higher quality of living standard. M-business and m-commerce may provide an opportunity in this regard. It is likely that such an ICT-enabled initiative could make a positive impact on the lives of the people and businesses in SSA. The diffusion of the basic device for m-business, namely, mobile telephones, could be interpreted to mean that the opportunity to deliver quality information in a real-time fashion is enhanced.

The Manobi projects in Senegal offer support for the preceding statement with concrete examples as to how such changes can be brought about. In the first phase of the Manobi project, farmers increased their returns by over 50% from a mobile phone-based market price system. The quality of lives (i.e., safety) of the Senegalese fishermen using the Manobi system was improved through the availability of up-to-date weather information. The fishermen also have the opportunity to increase their earning through quality information about market prices. Additionally, according to Manobi (2003), the project was able to persuade Sonatel to install a cell phone base station near the beach at Kayar. This benefits the fishermen as well as the neighboring communities.

There are, however, several challenges facing the emergence of m-business and m-commerce that policymakers in the region must address. A few of these are briefly discussed as follows.

Technology

Policymakers in the SSA region must continue to encourage the deregulation of the telecommunications industry. Studies show that teledensity (number of telephone users per 100) tends to be higher where governments are not involved in the provision of telephony to their citizens (Ifinedo, 2005a, 2005b). SSA countries wishing to adopt m-business must be aware of developments in the technology and must be willing to accommodate new concepts that may emerge with the initiative. Bluetooth is one example that is diffusing in the developed West which is relatively unknown in SSA. To ensure a smooth diffusion of m-business and m-commerce, governments in the region must invest in building related infrastructure. Telephony must be extended to both rural and urban centers in SSA to avoid the inadequacies that were seen with landlines, which tend to be concentrated around urban centers (Okoli, 2003; Ifinedo, 2005a). When m-business emerges fully in SSA, policymakers must not downplay security and legal issues often accompanying the use of such technologies. A lack of attention in such areas might negatively affect peoples' attitudes toward the initiative.

Illiteracy

This is a major challenge for any emerging m-business initiative in SSA. The region has the highest rate of illiteracy in the world (Ifinedo, 2005a; CIA, 2005). In order to fully reap the benefits of ICT products for improving economic conditions, efforts must be directed toward improving the literacy levels in SSA.

Costs

Efforts must focus on making mobile telephones (handsets) available to more people in SSA. The current statistic of nine mobile phone users per 100 persons could be improved.

Governments may consider subsidizing such facilities for poor farmers and artisans in underserved communities.

Health Concerns

Governments in SSA must follow developments regarding health concerns that have been linked to the use of mobile telephones (ARPNSA, 2005). For example, the preceding report states:

Concerns have been raised about the normal mobile phone, which has the antenna in the handset. In this case, the antenna is very close to the user's head during normal use of the telephone and there is concern about the level of microwave emissions to which the brain is being exposed.

Culture

Governments in the SSA region must begin to sensitize their populace about the benefits of using ICT products to improve their quality of life and living standards. Efforts must be directed at instilling a new way of thinking about m-business and related initiatives. Arguably, m-business might be facing an uphill task spreading in SSA if the cultural undertones are not conductive for digital transactions. Okoli (2003, p.16) comments that "Africans do not have the culture of buying a product without [having] tactile contacts." For m-business to thrive, a change of culture may be necessary.

REFERENCES

Annerose, D. (2001). *Using ICTs to increase incomes for farmers and fishermen in Senegal (Acacia II).* Retrieved March 7, 2006, from http://www.idrc.ca/wsis/ev-8117-201-1-DO_TOPIC.html

Anonymous. (2006). *M-commerce.* Retrieved April 5, 2006, from http://searchmobilecomputing.techtarget.com/sDefinition/0,,sid40_gci214590,00.html

ARPNSA (Australian Radiation Protection and Nuclear Safety Agency). (2005). *Mobile telephones and health effects.* Retrieved May 30, 2006, from http://www.arpansa.gov.au/pubs/factsheets/013.pdf

Avgerou, C. (1998). How can IT enable economic growth in developing countries. *Information Technology for Development, 8,* 15-28.

BBC News. (2002a). *Africans embrace mobiles and the Net.* Retrieved April 7, 2006, from http://news.bbc.co.uk/2/hi/technology/2290486.stm

BBC News. (2002b). *Mobiles find right price for farmers.* Retrieved April 7, 2006, from http://news.bbc.co.uk/2/hi/technology/2290540.stm

Castells, M. (1999). *Information technology, globalization and social development.* UNRISD Discussion Paper No. 114, United Nations, New York.

CIA. (2005). *World factbook: Country report (Nigeria).* Retrieved November 1, 2005, from http://www.cia.gov/cia/publications/

G8 DOT Force. (2001). *Issue objectives for the Genoa Summit Meeting 2001: DOT Force.* Retrieved December 12, 2005, from http://www.g8.utoronto.ca/

Ifinedo, P. (2005a, January 19-21). E-government initiative in a developing country: Strategies and implementation in Nigeria. *Proceedings of the 26th McMaster World Congress/6th World Congress on Electronic Business* (pp. 1-11), Hamilton, Ontario, Canada.

Ifinedo. (2005b). Measuring Africa's e-readiness in the global networked economy: A nine-country data analysis. *The International Journal of Education and Development using Information and Communication Technology, 1*(1), 53-71.

ITU. (2004). *Mobile cellular, subscribers per 100 people 2004.* Retrieved from http://www.itu.int/ITU-D/ict/statistics/at_glance/cellular04.pdf

Manobi. (2003). Retrieved February 3, 2006, from http://www.sustainableicts.org/infodev/Manobi.pdf

Marcussen, C.H. (2002). *Mobile data and m-commerce in Europe.* Retrieved April 3, 2006, from http://www.crt.dk/uk/staff/chm/wap/sms.pdf

McKinsey Quarterly. (2006). *M-commerce: Advantage, Europe.* Retrieved May 13, 2006, from http://www.mckinseyquarterly.com/article_abstract_visitor.aspx?ar=821&L2=22&L3=78

Okoli, C. (2003). *Expert assessments of e-commerce in Sub-Saharan Africa: A theoretical model of infrastructure and culture for doing business using the Internet.* Unpublished PhD thesis, Louisiana State University, USA.

Ololube, N. P., & Ifinedo, P. (2006). Photo of an African farmer in his field. Retrieved June 2, 2006, from http://www.ifinedo.com

Plant, S. (n.d.). *On the mobile: The effects of mobile telephones on social and individual life.* Retrieved April 8, 2006, from http://www.motorola.com/mot/doc/0/234_MotDoc.pdf

Sarker, S., & Wells, J.D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35-40.

Sadeh, N., & Sadeh, N. (2002). *M-commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

Singh, A. (2000). *Electronic commerce: Some implications for firms and workers in developing countries*. Discussion Paper, International Institute for Labour Studies (IILS), International Labour Organization. Retrieved June 2, 2004, from http://www.ilo.org/public/english/bureau/inst/

Turban, E., Lee, J., King, D., & Chung, H.M. (2000). *Electronic commerce: A managerial perspective.* Englewood Cliffs, NJ: Prentice-Hall.

UN ICT TASK Force. (2004). Retrieved January 8, 2005, from http://www.unicttaskforce.org/index.html

World Bank. (2001a). *World development report 1999/2000.* Retrieved February 3, 2004, from http://www.worldbank.org/wdr/2000/pdfs/engtable2.pdf

World Bank. (2001b). *Prospects for developing countries and world trade.* Retrieved February 2, 2004, from http://www.worldbank.org/prospects/gep2001/chapt1.pdf

WEF (World Economic Forum). (2003). Retrieved March 2, 2006, from http://www.weforum.org/pdf/Global_Competitiveness_Reports/Reports/GITR_2002_2003/ICT_Africa.pdf

Zwass, V. (1996). Electronic commerce: Structures and issues. *International Journal of Electronic Commerce, 1*(1), 3-23.

KEY TERMS

Bluetooth: This technology describes how mobile or cellular phones, PDAs, and computers can easily be interconnected using a short-range wireless connection.

Information and Communication Technology (ICT): Information systems and communication technologies that allow data to be processed, stored, shared, and transmitted over networks.

Mobile Business (M-Business): The sharing of business information, maintaining business relationships, and conducting business transactions using a wireless radio wave transmission technology.

Mobile Commerce (M-Commerce): The buying and selling of goods and services through a network of wireless radio wave transmission technology.

Mobile Telephone: A handheld device used in mobile telephony; it is also sometimes known as a handset.

Mobile Telephony: The use of wireless transmission technology to connect users.

Personal Digital Assistant (PDA): A handheld device that organizes personal data; includes an address book, task list, memo pad, and calculator, among others.

Short Message Service (SMS): A content delivery technology introduced in 1997 that permits the sending of short texts between users in a mobile telecommunications environment.

Sub-Saharan Africa (SSA): The region of Africa excluding the northern part of the continent and the Republic of South Africa.

Supply Chain Management: Management of the value-added chain, from the supplier of goods and services to the final consumer.

Mobile Television

Frank Hartung
Ericsson GmbH, Germany

Markus Kampmann
Ericsson GmbH, Germany

Uwe Horn
Ericsson GmbH, Germany

Jan Kritzner
Aachen University, Germany

INTRODUCTION

During the last 10 years, the TV landscape has changed quite significantly. With a relatively low investment, consumers can get access to several hundreds of TV channels. Free TV stations are trying to establish new revenue sources in order to become less dependent on revenues from advertisements. FreeTV and PayTV service providers operating via cable or satellite networks are faced with competition from IPTV service providers, which are offering video-on-demand services to broadband Internet users.

An important emerging trend is Mobile Television (MobileTV). The basic idea of delivering TV services to mobile users has been around for quite some time. Between 1982 and 1985 Seiko sold a wristwatch including a small TV screen, which had to be connected to an external receiver carried at the belt. The display size was very small, roughly half the size of a modern smartphone, and also the quality of the monochrome display was limited. Later, when LCD displays became available at cheap prices, several manufacturers started to build handheld TV receivers. However, this approach failed for a number of reasons: First the additional TV reception equipment was often too large and heavy to be carried around all the time. Secondly, it was often difficult to receive noise-free pictures, simply because terrestrial TV networks had not been engineered for portable reception.

Nowadays, mobile phones with multimedia capabilities and excellent display qualities have become commodities (Lee, Byun, Lee, & Kim, 2003; Koike, Matsumoto, & Kokubun, 2006), and third-generation (3G) cellular networks like UMTS provide sufficient capacity and data rates to deliver multimedia services to mobile users. Those networks also provide very good coverage, which gives mobile users access to MobileTV services wherever they are. Upcoming mobile broadband technologies like HSDPA will give further quality improvements due to increased data rates. With 3GPP MBMS and 3GPP2 BCMCS, broadcast/multicast extensions have been added to 3G standards and will become commercially available during 2007/2008 (Toenjes, 2004; Bakhuizen & Horn, 2005).

Apart from the cellular broadcast/multicast evolution, also digital TV transmission standards have produced variants addressing broadcast delivery to mobile devices (Weck & Wilson, 2006). DVB-T for instance has produced a variant called DVB-H, where the *H* stands for *handheld*. Even the digital transmission standard DAB, originally developed as a digital replacement for analog FM radio, developed a variant called T-DMB, which allows broadcasting of multimedia content to mobile devices.

This article gives an overview of the main technology and service trends in MobileTV. We start with an overview about existing MobileTV services. We then discuss the various transport options for delivering MobileTV services. At the end we go through some important service layer components.

MobileTV

MobileTV Services Today

During 2004 and 2005 many mobile operators launched commercial MobileTV services. In the simplest case, this is just a re-broadcast of existing TV channels.

In order to get access to a MobileTV offering, users have to pay a monthly fee. The basic package contains access to live TV channels; for on-demand premium content (for instance, sports channels), end users must pay separately.

However, existing TV channels are using formats that are not adapted to the behavior of mobile users. Studies have shown that mobile users consume MobileTV services in a very different way than fixed TV services (Södergard, 2003). Most noticeable, they spend less overall time per viewing session. The time spent per session is typically around 5

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

minutes in the mobile case. Likewise the time spent on individual programs is around 1-2 minutes. This behavior makes it necessary to develop new content formats, better tailored to the needs of mobile users. This was addressed during 2005 by some operators, which started to add mobile-specific content provided by existing TV channels to their offerings. An example is a 20-minute summary with the highlights of the week combined with the latest news which is looped over the day and updated once or twice per day.

The services provided by mobile operators today also show a clear trend towards combined offerings including not only live channels, but also personalized video-on-demand content.

Delivering MobileTV services to the mobile phone as an interactive communication device opens up a lot of opportunities for creating new services around the TV experience, and novel services emerge (Rauchenbach, 2005).

An example is interactive MobileTV. Already today there are many formats in traditional TV allowing users to interact with a TV show by sending back SMS for voting, chatting, purchase of ring tones, and so on. Those formats are often used by music stations, targeting younger people. On a mobile device, it is possible to seamlessly integrate the interaction and the program. For instance, a voting request can be presented as an interactive menu from which the user can easily make his selection. By browsing to the right answer and confirming the choice, voting services becomes much more convenient to use, compared to creating and sending an SMS which requires a lot of typing. This demands for advanced user interfaces (e.g., Knoche, 2005).

Systems for Delivering MobileTV Services

Commercial MobileTV services today are delivered over cellular networks. In particular, 3G cellular networks like UMTS provide sufficient capacity and data rates to achieve a good quality of experience. Those networks also provide very good coverage, which gives mobile users access to MobileTV services wherever they are. Upcoming mobile broadband technologies like HSDPA will give further quality improvements due to increased data rates.

Most of the existing MobileTV services are built upon the packet-switched streaming service (PSS) as it was standardized by 3GPP (3GPP TS 26.234, 2004; Elsen, Hartung, Horn, Kampmann, & Peters, 2001). The advantage of 3GPP PSS is its wide availability and support in existing mobile multimedia phones, and some effort has been made to bring PSS's services to unidirectional links (Yoshimura & Ohya, 2004). Since PSS uses a point-to-point connection between each client and a media server, it does not scale very well with an increasing number of simultaneous users.

Therefore, in 2003 3GPP and 3GPP2 started to address broadcast/multicast services in GSM/WCDMA and CDMA2000, respectively. In 3GPP the work item is called Multimedia Broadcast and Multicast Service (MBMS) (3GPP TS 23.246, n.d.; 3GPP TS 26.346, n.d.). In 3GPP2 it is called BroadCast and MultiCast Service (BCMCS). The specifications of cellular broadcast services were functionally frozen during 2004. 3GPP MBMS and 3GPP2 BCMCS have many commonalities. Both of them add the following capabilities to cellular networks:

- A set of functions that control the broadcast/multicast delivery service; MBMS uses the term Broadcast/Multicast Service Center, whereas in BCMCS it is called "BCMCS Controller."
- Broadcast/multicast routing of data flows in the core network.
- Efficient radio bearers for point-to-multipoint radio transmission within a cell.

Both MBMS and BCMCS are introducing only small changes to the existing radio and core network protocols. In the higher protocol layers, the same media codecs and associated transport protocols as for packet-switched streaming are used. This reduces the implementation costs both in terminals and in the network, and makes cellular broadcast a relatively cheap technology. Another advantage of cellular broadcast is that mobile operators can retain their established business models. Current services, such as MobileTV, will greatly benefit from the capacity-boosting effect of broadcast capabilities. Certainly cellular broadcast will also stimulate the development of new, mobile, mass-media services. Likewise, cellular broadcast will enable operators to provide a full triple-play service offering—telephony, Internet, and TV—for mobile handheld devices in a cost-effective way over a common service and network infrastructure.

Not only cellular networks have addressed mobile broadcast services, also digital TV transmission standards have produced several variants addressing mobile broadcast extensions.

One of them is DVB-H (Digital Video Broadcasting for Handheld—Faria, Henriksson, Stare, & Talmola 2006; ETSI EN 302 304, 2004), which can be regarded as an extension to DVB-T (Ladenbusch, 2006), an European standard for conventional terrestrial video services. DVB-H adds new features to the physical and link layer to reduce the power consumption in the receiver and to allow for a more robust transmission as it is needed for mobile devices. DVB-H may reuse the frequencies of old analogue television services, but competes with digital non-mobile television services for the spectrum.

T-DMB (terrestrial digital multimedia broadcasting) is a Korean extension of the digital audio broadcasting (DAB) standard for digital radio. Many countries allocated radio frequencies for DAB, but in most countries the commercial success was limited. This spectrum can now be reused by

T-DMB—that is, it could be deployed quickly. A companion standard S-DMB offers the same service over satellite links, and it is possible to use both DMB standards jointly. For example, the reception quality can be optimized by using S-DMB for large-area coverage and add T-DMB for improved indoor coverage.

T-DMB was called DMB-T for a long time. However, China decided to name its conventional digital TV standard DMB-T, and therefore the Korean standard was renamed. Though not specially tailored for mobile television, DMB-T is able to compete with other approaches for unidirectional transport.

First concepts are underway to integrate DVB-H, DMB, and 3G mobile networks under a common system concept. For example, DXB (digital extended broadcast) tries to establish a common IP-layer and common source coding recommendations for all mentioned transport options (Schäfer, 2003).

Common to non-cellular, broadcast-only systems is the lack of an additional interactive channel, which can be used for feedback or personalized services. This has been recognized as a potential drawback, and integration between broadcast-only and cellular networks has been proposed in the past (Tuttlebee, Babb, Irvine, Martinez, & Worrall, 2003; Pangalos, Chew, Aghvami, & Tafazolli, 2004). However, there are certain co-existence issues that need to be addressed. For instance, in Europe the upper part of the digital TV spectrum, which could be used for DVB-H, is close to the GSM900 uplink. This might lead to interference problems in a combined DVB-H/GMS900 system. Also the business model becomes more complex since an interactive broadcast service provider has to make business agreements with both the DVB-H and the cellular network operator.

MobileTV Service Layer Standardization

The existence of many similar yet different MobileTV standards has been recognized as a disadvantage, especially for service providers that want to distribute their services over different networks supporting different standards. The Open Mobile Alliance (OMA) has at least defined a service layer system called OMA mobile broadcast services, which allows service providers to better separate service generation and management from actual broadcast distribution.

Apart from media transport, which can be realized over different access systems based on already existing standards, OMA BCAST addresses more service-related components of MobileTV service. An example is the machine-readable metadata that is necessary to receive the service, and its transmission. Such metadata is called electronic service guide (ESG), and is for example used to convey information about upcoming services and shows. Interactivity, typically also an important component, enables interaction between the end user and the service provider and its service, enabling functionality like voting or spontaneous purchase of goods.

A very important area in OMA BCAST is content and service access protection. In a commercial MobileTV offering for which a mobile operator wants to charge its customers, it is required to restrict the access to only those services a user has subscribed to. Service access protection is achieved by encrypting the audiovisual data, using strong stream ciphers with adequate key lengths. The keys used to decrypt the data are usually conveyed in a key hierarchy. That means a shared secret key is somehow established between the transmitting server and receiving client. There are different ways of establishing this shared secret, and in fact the different service access protection systems differ most in the principle of how this is done. Changing long-term keys are then encrypted with the shared secret key and transmitted to the receiving client. Such keys do not change too often, maybe in the order of hours or days. They are not directly used to encrypt the media data; rather they are used to encrypt frequently changing short-term keys, which are broadcast together with the protected audiovisual data. Those short-term keys change often, for example in the order of seconds, to avoid that possibly compromised keys are published and used by unauthorized parties. The short-term keys are used to decrypt the audiovisual data. Thus, the key hierarchy consists of a shared secret, which is used to protect long-term keys, which are used to protect short-term keys, which are used to protect the audiovisual data. In 3G MBMS, a service access protection mechanism has been defined that uses the smartcard present in GSM and 3G mobile phones to establish a mutually trusted shared secret. In DVB-H, a service access protection mechanism has been defined that uses special pre-shared keys that are designed for secure content exchange using DRM (digital rights management) systems (Hartung, 2004). Another service access protection mechanism in DVB-H specifies just an interface to proprietary service access protection systems, a model that has long been used in legacy Pay TV systems.

FUTURE TRENDS

In order to attract more and more mobile users, new content formats, better tailored to the needs of mobile users and in particular produced for mobile channels, will evolve. The similar holds for advertisements. The interaction and mobile positioning capabilities will trigger many new personalized add-on services. For example, it is very likely that both the offered programs and the inserted advertisements will be tailored to the user and the context, like location. Further, services will evolve from pure audiovisual services with real-time video and audio tracks, to feature and media-rich services, containing for example embedded pictures, text pages, clips, executables, and graphical navigation possibilities, like in a computer game or interactive Web page today. Mass content and personalized content will exist side by side.

Aggregation of highly personalized channels, be it in the network or in the terminal, will soon complement the linear TV programs provided today. Content is aggregated based on personal preferences and no longer controlled by content aggregators like TV stations.

Also, TV services and communication services will converge, for example there might be multi-player shows with virtual teams of several players, all sitting at home, but virtually connected via their keyboards and voice connections. There will also be a convergence between fixed and mobile services.

In general, MobileTV services will become rather integrated and bundled with other services, like Internet access, telephone, videophone, and mobile communication services. In order to exploit synergies, it is very likely that these different services will use more and more common components, and they might for example all be based on IP-based protocols and technologies, irrespective of the underlying access technology.

CONCLUSION

Although TV is a service known for many years, delivering TV services to mobile devices is still a comparatively new area, and the design requirements are just emerging (Knoche & McCarthy, 2005). However, there is a strong end user interest in those services. The mobile distribution channel will open new revenue streams for TV content owners and service providers.

Commercial MobileTV services have already been launched in several cellular networks. For mobile operators, MobileTV could become the looked-for killer application, and upcoming broadcast extensions like 3GPP MBMS and 3GPP2 BCMCS will help to further increase the capacity of cellular networks for delivering mass content.

There are also non-cellular standards like DVB-H and T-DMB evolving from existing digital broadcast systems. However, those standards are pure broadcast distribution systems, lacking support for on-demand and personalized services. However, in order to provide more than plain TV services, they need to become tightly integrated with cellular networks, which could then be used for service interaction and delivering personalized content.

The development of a common service layer, which can be used to achieve a good integration between different access technologies, is addressed by the Open Mobile Alliance forum (OMA, 2005).

In the future we will see new content and advertisement formats, better tailored to mobile users. The interaction and mobile positioning capabilities will trigger many new personalized add-on services. Highly personalized channels, compiled in the network or in the terminal, will soon complement the linear TV programs provided today.

REFERENCES

Bakhuizen, M., & Horn, U. (2005). Mobile broadcast/multicast in mobile networks. *Ericsson Review, 1,* 6-13.

Elsen, I., Hartung, F., Horn, U., Kampmann, M., & Peters, L. (2001). Streaming technology in 3G mobile communication systems. *IEEE Computer, 34*(9), 46-53.

ETSI EN 302 304. (2004, November). *Transmission system for handheld terminals (DVB-H).*

Faria, G., Henriksson, J.A., Stare, E., & Talmola, P. (2006). DVB-H: Digital broadcast services to handheld devices. *Proceedings of the IEEE, 94*(1), 194-209.

Hartung, F. (2004). Mobile digital rights management. In E. Becker, W. Buhse, D. Günnewig, & N. Rump (Eds.), *Digital rights management—Technological, economic, legal and political aspects* (2nd ed.) (pp. 138-149). Berlin: Springer.

Koike, A., Matsumoto, S., & Kokubun, H. (2006). Personal mobile DTV cellular phone terminal developed for digital terrestrial broadcasting with Internet services. *Proceedings of the IEEE, 94*(1), 281-288.

Knoche, H. (2005). A user-centred mobile television consumption paradigm. *Proceedings of the 8th Human Centred Technology Postgraduate Workshop,* Sussex, UK.

Knoche, H., & McCarthy, J.D. (2005). Design requirements for mobile TV. *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services* (pp. 69-76). New York: ACM Press.

Ladebusch, U., & Liss, C. A. (2006). Terrestrial DVB (DVB-T): A broadcast technology for stationary portable and mobile use. *Proceedings of the IEEE, 94*(1), 183- 193.

Lee, J., Byun, S.-K., Lee, J.-D., & Kim, T.-Y. (2003). Evaluation of technological innovation in the cellular phone display. *Proceedings of the Portland International Conference on Management of Engineering and Technology* (pp. 140-149).

Pangalos, P., Chew, K. A., Aghvami, H., & Tafazolli, R. (2004, November 4-5). The mobile VCE architecture for the interworking of mobile and broadcast networks. *Proceedings of WWRF12,* Toronto, Canada.

Rauschenbach, U. (2005). Interactive TV meets mobile computing. In N. Davies, T. Kirste, & H. Schumann (Eds.), *Mobile Computing and Ambient Intelligence: The Challenge* *of Multimedia, Dagstuhl Seminar Proceedings,* Dagstuhl, Germany.

Schäfer, R. (2003, October). Digital Extended Broadcasting—DXB. *Proceedings of the Wiesbadener Medienkolloquium,* Wiesbaden, Germany.

Södergard, C. (2003). *Mobile television—Technology and user experiences.* Espoo, Finland: VTT.

3GPP TS 23.246. (2004). *Technical specification group services and system aspects; Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (release 6).* 3rd Generation Partnership Project.

3GPP TS 26.234. (2004). *Technical specification group services and system aspects; transparent end-to-end Packet-switched Streaming Service (PSS) (release 6).* 3rd Generation Partnership Project.

3GPP TS 26.346. (2004). *Technical specification group services and system aspects; Multimedia broadcast/multicast service (MBMS); protocols and codecs (release 6).* 3rd Generation Partnership Project.

Tönjes, R., Horn, U., & Hartung, F. (2004, November). Business and technology challenges for mobile broadcast services in 3G. *Proceedings of the 12th Wireless World Research Forum Meeting (WWRF)*, Toronto, Canada.

Tuttlebee, W., Babb, D., Irvine, J., Martinez, G., & Worrall, K. (2003). Broadcast and mobile communication: Interworking—Not convergence. *EBU Technical Review,* 1-11.

Weck, C., & Wilson, E. (2006). Broadcasting to handhelds—An overview of systems and services. *EBU Technical Review,* 1-11.

Yoshimura, T., & Ohya, T. (2004). Mobile broadcast streaming service and protocols on unidirectional radio channels. *IEICE Transactions on Communication, E87-B*(9), 2596-2604.

KEY TERMS

BroadCast and MultiCast Service (BCMS): Extension of 3G mobile networks by multicast/broadcast techniques. Standardized within 3GPP2.

Digital Multimedia Broadcasting (DMB): Extension of the Digital Audio Broadcasting (DAB) standard for TV content distribution. Variants for terrestrial distribution (T-DMB) and satellite distribution (S-DMB) exist. Mainly exploited in Korea.

Digital Video Broadcasting for Handheld (DVB-H): Extension to DVB-T for broadcasting of TV content to mobile terminals. Special techniques for low power consumption in the receiver are included.

Digital Extended Broadcast (DXB): System concept for an integrated mobile TV delivery solution for DMB, DVB-H, and 3G mobile networks.

High-Speed Downlink Packet Access (HSDPA): Packet-based data service in 3G networks with data transmission up to 8-10 Mbit/s.

Multimedia Broadcast and Multicast Service (MBMS): Extension of 3G mobile networks by multicast/broadcast techniques. Standardized within 3GPP.

Open Mobile Alliance Broadcast (OMA BCAST): Working group within the standardization body OMA defining the service layer system, "OMA Mobile Broadcast Services."

Packet-switched Streaming Service (PSS): 3GPP standard for streaming of multimedia content over mobile networks. Used for the distribution of MobileTV over 3G networks

3GPP (3rd Generation Partnership Project) /3GPP2: Standardization bodies for the specification of mobile networks, especially 3G mobile networks.

Uni-/Bidirectional: Unidirectional networks act like conventional broadcast stations where data can only be transmitted in one direction to the viewer. Bidirectional networks or links can transmit feedback from the user and allow for interactivity.

Uni-/Multicast: A unicast transmission is directly sent to one receiver, hence a TV program seen by n users is transmitted n times. In a multicast network data is transmitted to a group of receivers jointly.

Mobile Text Messaging Interface for Persons with Physical Disabilities

Cheng-Huei Yang
National Kaohsiung Marine University, Taiwan

Li-Yeh Chuang
I-Shou University, Taiwan

Cheng-Hong Yang
National Kaohsiung University of Applied Sciences, Taiwan

Jun-Yang Chang
National Kaohsiung University of Applied Sciences, Taiwan

INTRODUCTION

Morse code has been shown to be a valuable tool in assistive technology, augmentative and alternative communication, and rehabilitation for some people with various conditions such as spinal cord injuries, non-vocal quadriplegics, and visual or hearing impairments. In this article, a mobile phone human-interface system using a Morse code input device is designed and implemented for the person with disabilities to send/receive SMS (simple message service) messages or make/respond to a phone call. The proposed system is divided into three parts: input module, control module, and display module. The data format of the signal transmission between the proposed system and the communication devices is the PDU (protocol description unit) mode. Experimental results revealed that three participants with disabilities were able to operate the mobile phone through this human-interface after four weeks' practice.

BACKGROUND

A current trend in high-technology production is to develop adaptive tools for persons with disabilities to assist them with self-learning and personal development, and lead more independent lives. Among the various technological adaptive tools available, many are based on the adaptation of computer hardware and software. The areas of application for computers and these tools include training, teaching, learning, rehabilitation, communication, and adaptive design (Enders & Hall, 1990; McCormick, 1994; Bower et al., 1998; King, 1999).

Many adapted and alternative input methods now have been developed to allow users with physical disabilities to use a computer. These include modified direct selections (via mouth stick, head stick, splinted hand, etc.), scanning methods (row-column, linear, circular), and other ways of controlling a sequentially stepping selection cursor in an organized information matrix via a single switch (Anson, 1997). However, they were not designed for mobile phone devices. Computer input systems, which use Morse code via special software programs, hardware devices, and switches, are invaluable assets in assistive technology (AT), augmentative-alternative communication (AAC), rehabilitation, and education (Caves, 2000; Leonard, Romanowski, & Carroll, 1995; Shannon, Staewen, Miller, & Cohen, 1981; Thomas, 1981; French, Silverstein, & Siebens, 1986; Russel & Rego, 1998; Wyler & Ray, 1994). To date, more than 30 manufacturers/developers of Morse code input hardware or software for use in AAC and AT have been identified (Anson, 1997; Yang, 2000, 2001, 2003; Yang, Chuang, Yang, & Luo, 2002, 2003; see also http://www.uwec.edu/Academic/Outreach/Mores2000/morse2000.html). In this article, we adopt Morse code to be the communication method and present a human-interface for persons with physical disabilities.

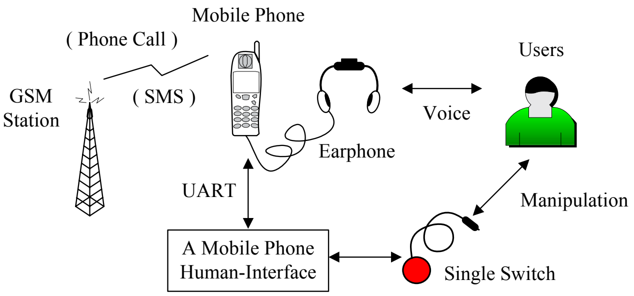
The technology employed in assistive devices has often lagged behind mainstream products. This is partly because the shelf-life of an assistive device is considerably longer then mainstream products such as mobile phones. In this study, we designed and implemented an easily operated mobile phone human-interface device by using Morse code as a communication adaptive device for users with physical disabilities. Experimental results showed that three participants with disabilities were able to operate the mobile phone through this human-interface after four weeks' practice.

SYSTEM DESIGN

Morse code is a simple, fast, and low-cost communication method composed of a series of dots, dashes, and intervals

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. System schematics of the mobile phone human-interface



in which each character entered can be translated into a predefined sequence of dots and dashes (the elements of Morse code). A dot is represented as a period ".", while a dash is represented as a hyphen, or minus sign, "-". Each element, dot or dash, is transmitted by sending a signal for a standard length of time. According to the definition of Morse code, the tone ratio for dot to dash must be 1:3. That means that if the duration of a dot is taken to be one unit, then that of a dash must be three units. In addition, the silent ratio for dot-dash space to character-space also has to be 1:3. In other words, the space between the elements of one character is one unit while the space between characters is three units (Yang et al., 2002).

In this article, the mobile phone human-interface system using a Morse code input device is schematically shown in Figure 1. When a user presses the Morse code input device, the signal is transmitted to the key scan circuit, which translates the incoming analog data into digital data. The digital data are then sent into the microprocessor, an 8051 single chip, for further processing. In this study, an ATMEL series 89C51 single chip has been adopted to handle the communication between the press-button processing and the communication devices. Even though the I/O memory capacity of the chip is small compared to a typical PC, it is sufficient to control the device. The 89C51 chip's internal serial communication function is used for data transmission and reception (Mackenzie, 1998). To achieve the data communication at both ends, the two pins, TxD and RxD, are connected to the TxD and RxD pins of an RS-232 connector. Then the two pins are connected to the RxD and TxD of an UART (universal asynchronous receiver transmitter) controller on the mobile phone device. Then, persons with physical disabilities can use this proposed communication aid system to connect their mobile communication equipment, such as mobile phones or GSM (global system for mobile communications) modems, and receive or send their messages (SMS, simple message service). If they wear an earphone, they might be able to dial or answer the phone. SMS is a protocol (GSM 03.40 and GSM 03.38) that was established by the ETSI (the European Telecommunications Standards Institute). The transmission model is divided into two models: text and PDU (protocol description unit). In this system, we use the PDU model to transmit and receive SMS information through the AT command of the application program (Pettersson, 2000). Structurally the mobile phone human-interface system is divided into three modules: the input module, the control module, and the display module. The interface framework is graphically shown in Figure 2. A detailed explanation is given as follows.

Input Module

A user's input will be digitized first, and then the converted results will be sent to the micro controller. The signal processing circuit can monitor all input from the input device, the Morse code. The results will be entered into the input data stream. When the user presses the input key, the micro-operating system detects new input data in the data stream, and then sends the corresponding characters to the display module. Some commands and/or keys, such as *OK, Cancel, Answer, Response, Send, Receive, Menu,* and *Exit,* have been customized and perform several new functions in order to accommodate the Morse code system. These key modifications facilitate the human-interface use for a person with disabilities.

Figure 2. Interface framework of mobile phone for persons with physical disabilities

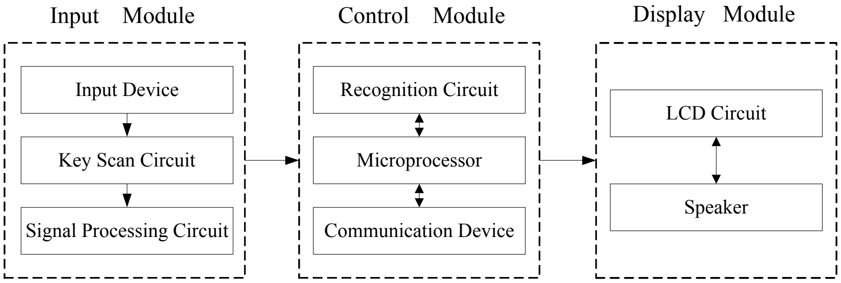
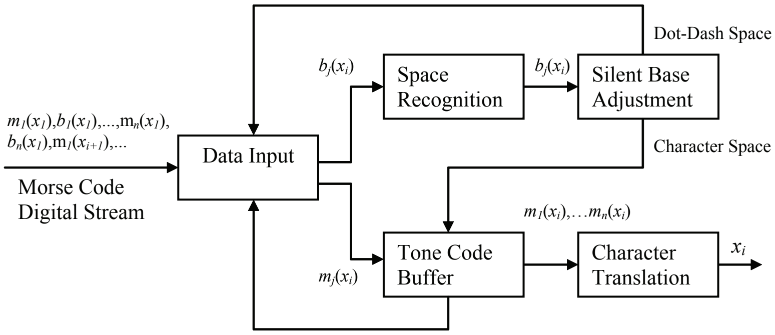


Figure 3. Block diagram of the Morse code recognition system



Control Module

The proposed recognition method is divided into three modules (see Figure 3): space recognition, adjustment processing, and character translation. Initially, the input data stream is sent individually to separate tone code buffer and space recognition processes, which are based on key-press (Morse code element) or key-release (space element). In the space recognition module, the space element value is recognized as a dot-dash space or a character space. The dot-dash space and character space represent the spaces existing between individual characters and within isolated elements of a character respectively. If a character space is identified, then the value(s) in the code buffer is (are) sent to character translation. To account for varying release speeds, the space element value has to be adjusted. The silent element value is sent into the silent base adjustment process. Afterwards, the character is identified in the character translation process.

A Morse code character, x_i, is represented as follows:

$$m_1(x_i),\ b_1(x_i),\ \ldots,\ m_j(x_i),\ b_j(x_i),\ \ldots,\ m_n(x_i),\ b_n(x_i)$$

where

$b_j(x_i)$: jth silent duration in the character x_i.

n: the total number of Morse code elements in the character x_i.

$m_j(x_i)$: the jth Morse code element of the input character x_i.

Display Module

Since users with disabilities have, in order to increase the convenience of user operations, more requirements for system interfaces than a normal person, the developed system shows selected items and system condition information on an electronic circuit platform, which is based on LCD (Liquid crystal display). The characteristics of the proposed system can be summarized as follows: (1) easy operation for users with physical disabilities with Morse code input system, (2) multiple operations due to the selection of different modes, (3) highly tolerant capability from adaptive algorithm recognition, and (4) system extension for customized functions.

RESULTS AND DISCUSSION

This system provides two easily operated modes, the phone panel and LCD panel control mode, which allow a user with disabilities easy manipulation. The following shows how the proposed system sends/receives an SMS message or makes/responds to a phone call.

SMS Receiving Operation

First, when users receive a message notification and want to look at the content, this system will provide phone panel and LCD panel control modes to choose from. In the phone panel mode, users can directly key-in Morse code "…" (as character 'S'). The interface system will go through the message recognition process, then exchange the message into the AT command "AT+CKPD='S', 1", to execute the "confirm" action of the mobile phone. The purpose of this process is the same as users keying-in "yes" on the mobile phone keyboard, then keying-in Morse code ". - - - ." (as key '↓'). The system will recognize the message, then automatically send the "AT+CKPD='↓', 1" instruction. The message cursor of the mobile phone is moved to the next line, or keys-in Morse code ". - . . -" (as key '↑') for moving it to the previous data line. Finally, if users want to exit and return to the previous screen, they only need to key-in Morse code ". . - ." (as character 'F'), and start the c key function on the mobile phone keyboard. If LCD panel mode is selected, one can directly follow the selected items on the LCD crystal, to execute the reception and message reading process.

SMS Transmitting Operation

Message transmission services are provided in two modes: phone panel and LCD panel. In the phone panel mode, continually type two times the Morse code ". - . - ." (as key '→'). The system will be converted into AT Command and transferred into mobile phone to show the selection screen of the message functions. Then continuing to key-in three times the Morse code ". . ." (as character 'S'), one can get into the editing screen of message content, and wait for users to input the message text data and receiver's phone number. The phonebook function can be used to directly save the receiver's phone number. After the input, press the 'yes' key to confirm that the message sending process has been completed. In addition, if the LCD panel mode is selected, one can follow the LCD selection prompt to input the service selection of all the action integrated in the LCD panel. Then the user goes through the interface and translates to a series of AT command orders, and batch transfers these into the mobile phone to achieve the control purpose.

The selection command "Answer a phone" displays on the menu of the LCD screen and can be constructed using Morse code. The participants could press and release the switch and input the number code ". - - - -" (as character '1') or hot key ". -" (as character 'A'). The mobile phone is then answered automatically. Problems with this training, according to participants, are that the end result is limited typing speed and users must remember all the Morse code sets of commands.

Three test participants were chosen to investigate the efficiency of the proposed system after practicing on this system for four weeks. Participant 1 (P1) was a 14-year-old male adolescent who has been diagnosed with cerebral palsy. Participant 2 (P2) was a 14-year-old female adolescent with cerebral palsy, athetoid type, who experiences involuntary movements of all her limbs. Participant 3 (P3) was a 40-year-old male adult, with a spinal cord injury and incomplete quadriparesis due to an accident. These three test participants with physical impairments were able to make/respond to phone calls or send/receive SMS messages after practice with the proposed system.

FUTURE TRENDS

In the future, a Morse code input device could be adapted to several environmental control devices, which would facilitate the use of everyday appliances for people with physical disabilities considerably.

CONCLUSION

To help some persons with disabilities such as amyotrophic lateral sclerosis, multiple sclerosis, muscular dystrophy, and other conditions that worsen with time and cause the user's abilities to write, type, and speak to be progressively lost, requires an assistive tool for purposes of augmentative and alternative communication in their daily lives. This article presents a human-interface for mobile phone devices using Morse code as an adapted access communication tool. This system provides phone panel and LCD panel control modes to help users with a disability with operation. Experimental results revealed that three physically impaired users were able to make/respond to phone calls or send/receive SMS messages after only four weeks' practice with the proposed system.

ACKNOWLEDGMENTS

This research was supported by the National Science Council, R.O.C., under grant NSC 91-2213-E-151-016.

REFERENCES

Anson, D. (1997). *Alternative computer access: A guide to selection.* Philadelphia: F. A. Davis.

Bower, R. et al. (Eds.). (1998). *The trace resource book—assistive technology for communication, control, and computer access.* Trace Research & Development Center, Waisman Center, University of Wisconsin – Madison, USA.

Caves, K. (2000). Morse code on a computer—really? *Proceedings of the 1st Morse 2000 World Conference,* Minneapolis, MN.

Enders, A., & Hall, M. (Ed.). (1990). *Assistive technology sourcebook.* Arlington, VA: RESNA Press.

French, J. J., Silverstein, F., & Siebens, A. A. (1986). An inexpensive computer based Morse code system. *Proceedings of the RESNA 9th Annual Conference* (pp. 259-261). Minneapolis, MN.

King, T. W. (1999). *Modern Morse code in rehabilitation and education.* Boston: Allyn and Bacon.

Leonard, S., Romanowski, J., & Carroll, C. (1995). Morse code as a writing method for school students. *Morsels, University of Wisconsin-Eau Claire, 1*(2), 1.

Mackenzie, I. S. (1998). *The 89C51 microcontroller* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

McCormick, J. A. (1994). *Computers and the Americans with Disabilities Act: A manager's guide.* Blue Ridge Summit, PA: Wincrest/McGraw-Hill.

Pettersson, L. (2000). *Dreamfabric.* Retrieved from http://www.dreamfabric.com/sms

Russel, M., & Rego, R. (1998). A Morse code communication device for the deaf-blind individual. *Proceedings of ICAART* (pp. 52-53). Montreal, Canada.

Shannon, D. A., Staewen, W. S., Miller, J. T., & Cohen, B. S. (1981). Morse code controlled computer aid for the nonvocal quadriplegic. *Medical Instrumentation, 15*(5), 341-343.

Thomas, A. (1981). Communication devices for the non-vocal disabled. *Computer, 14,* 25-30.

Wyler, A. R., & Ray, M. W. (1994). Aphasia for Morse code. *Brain and Language, 27*(2), 195-198.

Yang, C.-H. (2000). Adaptive Morse code communication system for severely disabled individuals. *Medical Engineering & Physics, 22*(1), 59-66.

Yang, C.-H. (2001). Morse code recognition using learning vector quantization for persons with physical disabilities. *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences, E84-A*(1), 356-362.

Yang, C.-H. (2003). An interactive Morse code emulation management system. *Computer & Mathematics with Applications, 46,* 479-492.

Yang, C.-H., Chuang, L.-Y., Yang, C.-H., & Luo, C.-H. (2002). An Internet access device for physically impaired users of Chanjei Morse code. *Journal of Chinese Institute of Engineers, 25*(3), 363-369.

Yang, C.-H., Chuang, L.-Y., Yang, C.-H., & Luo, C.-H. (2003). Morse code application for wireless environmental control system for severely disabled individuals. *IEEE Transactions on Neural System and Rehabilitation Engineering, 11*(4), 463-469.

KEY TERMS

Adaptive Signal Processing: The processing, amplification, and interpretation of signals that change over time through a process that adapts to a change in the input signal.

Assistive Technology (AT): A generic term for a device that helps a person accomplish a task. It includes assistive, adaptive, and rehabilitative devices, and grants a greater degree of independence to people with disabilities by letting them perform tasks they would otherwise be unable to perform.

Augmentative and Alternative Communication (AAC): Support for and/or replacement of natural speaking, writing, typing, and telecommunications capabilities that do not fully meet a communicator's needs. AAC, a subset of AT (see above), is a field of academic study and clinical practice, combining the expertise of many professions. AAC may include unaided and aided approaches.

Global System for Mobile Communications (GSM): The most popular standard for global mobile phone communication. Both its signal and speech channels are digital, and it is therefore considered a second-generation mobile phone system.

Morse Code: A transmission method implemented by using just a single switch. The tone ratio (dot to dash) in Morse code has to be 1:3 per definition. This means that the duration of a dash is required to be three times that of a dot. In addition, the silent ratio (dot-space to character-space) also has to be 1:3.

Simple Message Service (SMS): A service available on digital mobile phones which permits the sending of simple messages between mobile phones.

Mobile Users in Smart Spaces

Loreno Oliveira
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

Constant technological advances are making *pervasive computing* (Weiser, 1991) a reality. Such advances have been enabling the rise of devices increasingly smaller, with larger storage space, novel wireless interfaces, and lower battery consumption. These innovative technologies are contributing to the emergence of a new sort of personal portable device, as well as a large number of sensors and actuators. Sensors and actuators are embedded into objects spread across the environment, while portable devices quietly inform such environments how users wish to interact with them. Therefore, personal mobile devices stand out currently as the interface between people and smart spaces.

A basic requirement in the context of pervasive computing is to allow users to access services seamlessly as they move across environments. This requirement demands from the underlying infrastructure the ability to transfer user sessions among access points (*handoff*), which is a well-known concern in the context of pervasive computing (Cui, Nahrstedt, & Xu, 2004; Banerjee, Das, & Acharya, 2005). Nevertheless, the effective delivery of services in smart spaces requires conceiving mechanisms for handling localized scalability, availability, and redundancy of services; load balancing among providers; and on-demand content transformation for different devices (Satyanarayanan, 2001; Raatikainen, Christensen, & Nakajima, 2002; Raman et al., 2002), henceforth *QoS issues*. These requirements rise as fundamental for promoting transparency and invisibility to the service usage, as well as delivering some level of QoS and optimized resource utilization (Kalasapur, Kumar, & Shirazi, 2006).

Currently, there are still no efforts for conceiving solutions to provide ubiquitous access and seamless usage of services while taking into account QoS issues. In this context, we define the dynamic provision of services as a set of requirements relevant to the seamless provision of services for mobile users plus mechanisms for dealing with QoS issues.

In this article we define and present the basis of our work about dynamical services provisioning for mobile users in smart spaces. We present an overview about our envisioned service provision infrastructure, as well as the main research challenges related to it. Finally, we discuss issues related to the current state of our research and point out research directions in this field.

BACKGROUND

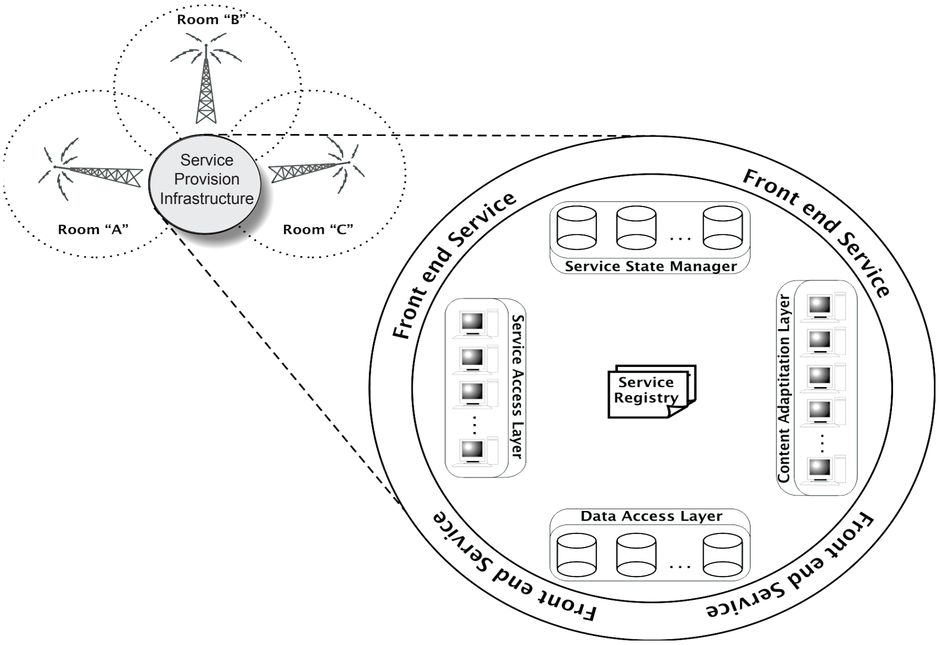
The interaction between users' devices and smart spaces occurs primarily through services advertised in those environments. The *service-oriented paradigm* (Papazoglou & Georgakopoulos, 2003) is especially suitable due to the dynamics of smart spaces, where resources may exist anywhere and applications running on mobile devices must be able to find out and use them at runtime.

In the context of smart spaces, user mobility is the main cause of such disturbances, but other factors may also cause temporary unavailability or degradation of services, for example, crash failures in the service providers, temporary network congestion, or peaks of overload on service providers. Smart spaces are primarily service-oriented environments, where part of the application logic is at the client side (e.g., in the form of helper applications for user profiles) and part at the server side (e.g., in the form of services offering some extra features for the users). When losing network connection, client applications also lose part of their capabilities, which are deployed as services in these environments. This scenario illustrates the first requirement of the applications aimed at smart spaces: the ability to switch between connected and disconnected modes. Applications and service implementations may utilize this connectionless period for performing some sort of preparation for when the connectivity is reestablished (Sairamesh, Goh, Stanoi, Padmanabhan, & Li, 2004).

When able to cope with off-line operations, both applications and the underlying infrastructure need to provide means for bringing transparency to the service usage, even when users move across access points. Sessions specify which client is using which services in a given access point.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Conceptual infrastructure for dynamic provision of services



When users move across access points, sessions need to be transferred between them, service providers need to be contacted, and the use of services needs to be reestablished from the point at which they were interrupted due to the change of location (handoff).

Currently there are several alternatives for dealing with application mobility, such as mobile IP (in short MIP), hierarchical MIP, cellular IP, teleMIP, and dynamic mobile agent (Saha, Mukherjee, Misra, Chakraborty, & Subhash, 2004). All these alternatives are suitable for IP-based applications and are embedded in the transport layer of the IP protocol stack. The major issue of these solutions is their unsuitability for dealing with *real-time traffic* management (Saha et al., 2004). In the context of smart spaces, this is an important limitation since these environments have many multimedia services, such as video and/or audio streaming services. New mechanisms need to be developed for providing multimedia services for mobile users (Cui et al., 2004; Banerjee et al., 2005).

Nevertheless, a great many issues may affect the quality of the provided service, such as adaptability to changes in the execution and communication environments, efficient use of communication resources, and high *availability* and stringent *fault-tolerance* (Raatikainen et al., 2002). The requirements for data accessed by these applications are quite similar. The underlying infrastructure must provide consistent, efficiently accessible, reliable, and a highly available information base (Raatikainen et al., 2002).

Presently, there are many approaches to deal with QoS issues in the context of smart spaces (Kumar & Song, 2005; Panagiotakis et al., 2003; Hingne, Joshi, Finin, Kargupta, & Houstis, 2003). However, current efforts focus on the implementation of mechanisms for only coping with QoS issues in isolation, that is, they do not provide any support for handoff. In the same manner, current solutions for dealing with handoff focus only on the handoff mechanisms and neglect other aspects that also impact the quality of the provided services.

The fact that distinct solutions regarding handoff management and *QoS assurance* for service provision are available does not imply that it is possible to aggregate these two features into one unique solution. Bridging two distinct solutions may be prohibitively burdensome, may introduce bugs and/or restrict interaction between the two software modules, and after that may also perform poorly because the two software modules were not designed for working together. Therefore, designing and implementing infrastructures able to not only deal with handoff, but also capable of delivering extra levels of QoS, is certainly the next step towards conceiving service provision infrastructures that better approximate from the pervasive computing principle of invisibility (Satyanarayanan, 2001).

DYNAMIC PROVISION OF SERVICES

We define dynamic provision of services as a set of requirements relevant to the seamless provision of services for mobile users plus mechanisms for dealing with QoS issues. The goal is to conceive an infrastructure for dealing

Table 1. QoS features

Feature	Expected Behavior
Seamless handoff scheme for both kinds of applications	**When moving across access points, users consuming both RPC and stream-based services will not experience interruptions at service usage, of course, only if users experience short times of disconnection.**
Data throughput	**Data throughput is granted, through resource reservations, for different kinds of data transfer, such as streamed video or file transfers.**
Availability of data and services	**We achieve better availability of services and data through redundancy, performed by the service and data access layers.**
Dynamic content adaptation	**Data transformation is performed on-demand in order to accomplish both static and dynamic requirements. For instance, the data throughput requirement may trigger a content adaptation process according to current network conditions.**
Policies for data delivery	**Users may specify which policies the underlying infrastructure must assume when necessary. For instance, users may wish to automatically reduce streamed video quality if the network gets congested instead of experiencing freezing in playback.**

with both handoff management and availability issues of computational resources at the server side. In the next sections we present a conceptual architecture of our envisioned infrastructure, along with some discussions regarding open issues and challenging aspects of our work.

Infrastructure Overview

In Figure 1 we present the conceptual architecture of our vision of a dynamic service provision infrastructure.

Client applications interact with the service provision infrastructure through a front end service. The front end service publishes an interface through which client applications can search/select/use services, and specifies QoS constraints of the required services. In Table 1 we briefly depict the QoS features we intend to address.

The front end service is also responsible for the authentication/authorization of the users. During this process, the front end service can also obtain information pertinent to eventual opened sessions and may use this information to execute part of the handoff procedure. Other relevant information can also be gathered during the authentication/authorization process, such as client profiles and device features (e.g., memory, CPU speed, display size/colors, and network technology). Some kind of network evaluation process can be coupled with the data exchange of the authentication/authorization process, so that instantaneous information regarding delay, bandwidth, and jitter may be identified in advance. This information, along with user profiles, device features, and defined QoS constraints, is used for applying on-demand data transformations, making the service usage as close as possible to clients expectations.

All complexities behind locating and controlling QoS parameters are tasks of the front end service and remain modules of the infrastructure. The front end service queries the service registry about the existence of providers for the needed services. The service registry can return a service handle or an error message. The error message can assume two meanings: there is no provider for a needed service or QoS constraints cannot be assured. Since a service handle is returned, the front end service uses it for contacting the service provider.

The access to service providers is mediated by a service access layer. The role of this layer is managing computational resources (essentially sets of computers), allocating and freeing them as the current workloads demand and/or permit. The service access layer also controls service instances, creating and destroying them dynamically according to the demand of such services and availability of computational resources. Therefore, the service access layer is also responsible to distribute workload among service providers.

In order to accept client requests, the service access layer must negotiate with the data access layer access to needed data according to client-defined QoS constraints. Just like the service access layer, the data access layer arbitrates and manages the access to data providers. It must control workload balancing among data providers and manage data copying/moving among providers. New resources can be allocated or freed as the current load conditions demand and/or permit.

The service state manager is only a symbolic entity. It denotes the need of providing means for reestablishing services interrupted due to user mobility. For instance, most RPC-based applications can be handled in a mobile environment through transport layer solutions, such as MIP. In this case, the "service state manager" can be mapped into the network elements required by MIP. On the other hand, stream-based services require some kind of collaborative cache scheme, where local buffers in the client side work together with cached data in the server side in order to deliver transparency to the handoff process.

The last major role of our infrastructure is the on-demand content adaptation. Client devices may drastically range in aspects such as processing power, storage space, user interface, and network connectivity, among others. These are some examples of static features that require some level of personalized content adaptation. Moreover, some dynamic features, such as network congestions, also demand personalized content. Dealing with such distinct constraints, both static and dynamic, requires delivering content using protocols and data formats best suited for each kind of device, according to their specific constraints and needs.

Technical Issues

A number of issues are identified in the context of the discussed infrastructure. Below we discuss some of them.

- **QoS Constraints:** Modeling user behaviors, computational resources, and expressing QoS constraints is yet an open issue in our work. Alternatives range from mathematical models, using matrixes and systems of equations (Sauvé et al., 2006), to queue systems along with simulation tools (Urgaonkar, Pacifici, Shenoy, Spreitzer, & Tantawi, 2005).
- **Security:** Normally, mobile devices are not shared among different users, and this property is used for recognizing users based on their portable devices. This authentication proceeding could raise problems, for example in the case of device thefts (Tatli, Stegemann, & Lucks, 2005). Security issues also cover data propagation over wireless medium (Grosche & Knospe, 2002) and authorization/authentication procedures in order to avoid opening the system ports for untrusted parts.
- **Creating and Destroying Service Instances on Demand:** Managing service instances can be split into two problems: how to develop services and how to monitor them. These problems create a set of questions that must be answered, such as: What are the requirements for writing services? What are their limitations? How do we deal with different underlying architectures? Does the service middleware have expressiveness enough for addressing different kinds of applications? What is the impact of the monitoring protocol over resource utilization? How do we estimate threshold values? Are static triggers enough? What happens with client requests interrupted due to server failures?
- **Access Layers:** Access layers encapsulate the access to sets of computational resources, acting as both resource schedulers and monitors of current demand vs. remaining computational power. If they are centralized entities, they become single points of failures, and if they fail, the access to all computational resources is lost. On the other hand, how do we contact them if they are distributed entities? What are the drawbacks of using hardcoded references? These remarks are also applicable to the front end service, which has the same properties as the access layers.
- **Distributed Databanks:** Conceiving distributed databanks is not a novel issue but is still a challenging task. Challenging issues include: data fragmentation and distribution, query processing and optimization, distributed concurrency control, reliability and commit protocols, and replicated data management. Techniques and algorithms for managing replicated data is also a concern, such as dealing with clocks, deadlock detection, and mutual exclusion.
- **Different Handoff Schemes:** User mobility may manifest itself in different faces and thus different kinds of handoffs. For instance, host-level (both user and device move) and user-level (only the user moves) handoffs concern how to identify the occurrence of handoff. Horizontal (handoff without changing in the underlying network technology) and vertical (handoff with changing in the underlying network technology) handoffs concern eventual changes in the underlying network technology. Providing transparency to the service usage requires definition of which kinds of mobility the underlying infrastructure will support.
- **Dynamic Content Adaptation:** Developing mechanisms for delivering personalized content for different kinds of applications—that is, different data formats—is an important open issue. On-demand data transformation may require considerable processing power, which implies distributed computation. This can be a challenging task since not all data formats can be processed in a parallel fashion (e.g., video rendering). Dynamic content adaptation may also require stringent integration with other software modules. For instance, when consuming streamed data, a cache adjustment may be necessary for hiding the extra time used for transforming data.

FUTURE TRENDS

The idea of dynamic service provision is still in its early days, and there is a lot of work still needed to implement a software infrastructure to promote such a scenario. The definition of which QoS parameters must be added or removed to the ones presented before is still source of discussion.

Since the requirements are well bounded, the next steps include directing efforts towards the issues depicted in the previous section. Only when we have a complete knowledge of the mechanisms we need for solving these issues will we be able to design and implement a prototype system.

CONCLUSION

In this article we presented our vision of a dynamic service provision infrastructure aimed at users of mobile devices. A high-level overview of our envisioned infrastructure was presented, including depiction and discussion of its major modules. We presented the main open issues and challenging tasks regarding the implementation of the introduced infrastructure.

We believe that the major contribution of this work is to make available a first discussion about all issues related to the subject. Our first efforts to specify such infrastructure, and the related open issues and challenging tasks, can serve as guidelines for research groups planning to propose new solutions in this field.

REFERENCES

Banerjee, N., Das, S. K., & Acharya, A. (2005). SIP-based mobility architecture for next generation wireless networks. *Proceedings of the 3rd International Conference on Pervasive Computing and Communications (PERCOM'05)* (pp. 181-190).

Cui, Y., Nahrstedt, K., & Xu, D. (2004). Seamless user-level handoff in ubiquitous multimedia service delivery. *Multimedia Tools Applications, 22*(2), 137-170.

Grosche, S.S., & Knospe, H. (2002). Secure mobile commerce. *Electronics & Communication Engineering Journal, 14*(5), 228-238.

Hingne, V., Joshi, A., Finin, T., Kargupta, H., & Houstis, E. (2003). Towards a pervasive grid. *Proceedings of the 17th International Parallel and Distributed Processing Symposium (IPDPS'03)* (p. 207.2).

Kalasapur, S., Kumar, M., & Shirazi, B. (2006). Evaluating service oriented architectures (SOA) in pervasive computing. *Proceedings of the 4th IEEE International Conference on Pervasive Computing and Communications (PERCOM'06)* (pp. 276-285).

Kumar, R., & Song, X. (2005). *GridLite: A framework for managing and provisioning services on grid-enabled resource limited devices.* Technical Report, Mobile and Media Systems Laboratory, HP Laboratories, Palo Alto, CA.

Panagiotakis, S., Koutsopoulou, M., Alonistioti, A., Houssos, N., Gazis, V., & Merakos, V. (2003). An advanced service provision framework for reconfigurable mobile networks. *International Journal of Mobile Communications, 1*(4), 425-438.

Papazoglou, M. P., & Georgakopoulos, D. (2003). Service-oriented computing: Introduction. *Communications of the ACM, 46*(10), 24-28.

Raatikainen, K., Christensen, H. B., & Nakajima, T. (2002). Application requirements for middleware for mobile and pervasive systems. *ACM SIGMOBILE Mobile Computing and Communications Review, 6*(4), 16-24.

Raman, B., Agarwal, S., Chen, Y., Caesar, M., Cui, W., Johansson, P., et al. (2002). The SAHARA model for service composition across multiple providers. *Proceedings of the 1st International Conference on Pervasive Computing (PERVASIVE'2002)* (pp. 1-14).

Saha, D., Mukherjee, A., Misra, I. S., Chakraborty, M., & Subhash, N. (2004). Mobility support in IP: A survey of related protocols. *IEEE Network, 18*(6), 34-40.

Sairamesh, J., Goh, S., Stanoi, I., Padmanabhan, S., & Li, C. S. (2004). Disconnected processes, mechanisms and architecture for mobile e-business. *Mobile Networks and Applications, 9*(6), 651-662.

Satyanarayanan, M. (2001). Pervasive computing: vision and challenges. *IEEE Personal Communications, 8*(4), 10-17.

Sauvé, J. P., Marques, F. T., Moura, J. A., Sampaio, M. C., Jornada, J., & Radziuk, E. (2006). Optimal design of e-commerce site infrastructure from a business perspective. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)* (pp. 178.3-178.3).

Tatli, E.I., Stegemann, D., & Lucks, S. (2005). Security challenges of location-aware mobile business. *Proceedings of the 2nd IEEE International Workshop on Mobile Commerce and Services (WMCS'05)* (pp. 84-95).

Urgaonkar, B., Pacifici, G., Shenoy, P., Spreitzer, M., & Tantawi, A. (2005). An analytical model for multi-tier Internet services and its applications. *Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'05)* (pp. 291-302).

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 66-75.

KEY TERMS

Helper Application: Small software deployed on resource-constrained devices. Helper applications implement only small sets of functionalities. They combine resources of other computers (e.g., advertised as services) to provide more powerful functionalities.

Portable/Mobile Device: Any low-sized device used to interact with other small devices and resources from smart spaces. Examples of portable/mobile devices are cellular phones, smart phones, PDAs, notebooks, and tablet PCs.

Proxy: A network entity that acts on behalf of another entity. Proxies' roles vary since data relays to the provision of value-added services, such as on-demand data adaptation.

Real-Time Traffic: Any data flow with stringent restrictions of throughput, such as maximum delay and jitter, and minimum bandwidth. Examples of real-time traffic are video broadcast or audio streaming.

Service-Oriented Computing (SOC): Distributed computing paradigm whose building blocks are distributed services. SOC stands out as the effective choice for advertising services to mobile devices in smart spaces.

User Profile: Set of information regarding user preferences when interacting with computational systems. Possible information of a profile of smart space users includes commonly used services and QoS constraints for such services.

Workload: Abstract notation for expressing the amount of work being performed by computational resources. Workloads indicate, for example, the number of requests being served by a certain host at a certain moment, which peripherals they are using, and how much CPU and memory they demand.

Mobile Video Transcoding Approaches and Challenges

Ashraf M. A. Ahmad
National Chiao Tung University, Taiwan

INTRODUCTION

Mobile access to multimedia contents requires video transcoding functionality at the edge of the mobile network for interworking with heterogeneous networks and services. Under certain conditions, the bandwidth of a coded video stream needs to be drastically reduced in response to changes in a highly constrained transmission channel, such as mobile.

Therefore, to guarantee quality of service (QoS) delivered to the mobile user, a robust and efficient transcoding scheme should be deployed in a mobile multimedia transporting network. In this article, we review several typical video transcoding architectures and major applications of video transcoding. We identify issues involved in accessing video streams through handheld devices and wireless networks. This article examines the challenges and limitations that face video transcoding schemes in a mobile multimedia transporting network. Then we explore different approaches for video transcoding schemes in a mobile multimedia transporting network.

BACKGROUND AND RELATED WORK

Current advances in mobile communications and portable client devices enable us to access multimedia content universally. However, when multimedia content becomes richer, including video and audio, it is difficult for wireless access because of many restrictions. On one hand, wireless connections usually have a lower bandwidth compared to wired ones and communication conditions change dynamically due to the effect of fading. On the other hand, portable client devices only have limited computing and display capabilities, which are not suitable for high-quality video decoding and displaying.

Concerning the heterogeneity issue, the previous era has seen a variety of developments in the area of multimedia representation and communication. In particular, we are beginning to see delivery of all types of multimedia data for all types of users in all types of conditions. In a diverse and heterogeneous world, the delivery path for multimedia content to a multimedia terminal is not straightforward, especially in the mobile communication environment. Access networks are various in nature, sometimes limited, and differ in performance. The characteristics of end user devices vary increasingly, in terms of storage, processing capabilities, and display qualities, as well as the natural environment (e.g., position, elucidation, temperature, changes). Finally, users are different by nature, showing dissimilar preferences, special usage, disabilities, and so forth.

The advance of multimedia systems has had a major influence in the area of image and video coding. The problem of interactivity and integration of video data with computer, cellular, and television systems is relatively new and subject to a great deal of research worldwide. As the number of networks, types of devices, and content representation formats increase, interoperability between different systems and different networks is becoming more important. Thus, devices such as gateways, multipoint control units, and servers must be developed to provide a seamless interaction between content creation and use.

The transporting of multimedia over wireless channels to mobile users is becoming a research topic of rapidly growing interest (Han et al., 1998; Warabino, Ota, Morikawa, & Ohashi, 2000; Mitchell, Pennebaker, Fogg, Chad, & LeGall, 1996; Shanableh & Ghanbari, 2000; Correia, Faria, & Assuncao, 2001). With the emergence of small wireless handset devices such as PDAs, video mobile, and so forth, it is expected that interactive multimedia will be a major source of traffic to these handset devices. These devices could be carried by users inside buildings when they are connected by a wireless local area network (LAN) or in vehicles when they will be connected to the cellular network, such as GPRS (Eleftheriadis & Anastassiou, 1995; Keesman, 1996). Wideband mobile communication systems such as IMT-2000 have also emerged, and there should be a mechanism to cope with a variety of media such as video provided to a mobile terminal.

Wireless transmissions use radio channels as the transmission media. Generally, radio links connect users to base stations which are connected to routers using wired links. The wireless segment "cells" provide mobility to a user while using the network. In contrast to wireline transmission links where the bandwidth can be easily increased and the channel quality can be guaranteed, the bandwidth of a wireless channel is limited because of spectrum allocation and physical limitations. The transmission quality of radio is easily affected by environments such as buildings, moving

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

objects, and atmosphere, as well as shielded obstacles and so forth. Moreover, because of the mobile nature of the users, the access point of a mobile user changes continuously. All these factors in wireless networks give rise to issues such as effective bandwidth allocation, high channel bit error rate, and user handover.

Moreover, because of its high traffic characteristics such as high bit rate, video will be the dominant traffic in multimedia streams and hence needs to be managed efficiently. Obviously for efficient utilization of network resources, video must be compressed to reduce its bandwidth requirement. Although there exist several compression techniques, MPEG [1, 2, and 4] is one of the most widely used compression algorithms for networked video applications. A wireless handset device, for instance a personal data assistant, can integrate voice, video, and data in one device. In contrast to solely text information, multimedia data can tolerate a certain level of error and fading. Therefore, although a wireless network has a high bit error rate when compared to a wireline network, it is possible to cost effectively transmit multimedia over wireless networks with acceptable quality.

For instance, the MPEG-2 compressed digital video content is being used in a number of products, including DVDs, camcorders, digital TVs, and HDTVs. In addition, tons of MPEG2 data have been stored already in different accessible multimedia servers. The ability to access this widely available MPEG-2 content on low-power end user devices such as PDAs and mobile phones depends on effective techniques for transcoding the MPEG-2 content to a more appropriate, low bit rate video.

Therefore mobile access to multimedia contents requires video transcoding functionality at the edge of the mobile network for interworking with heterogeneous networks and services, changing bit rates, and so forth. This transcoding mechanism should tickle the aforementioned issues in transmitting video in a mobile and wireless network.

FUNCTIONS OF TRANSCODING TECHNIQUES

Building a good video transcoding for mobile devices poses many challenges. To meet these challenges, a various kind of transcoding function is provided. This paragraph will describe these functions in detail.

The first function is bit rate adaptation. Bit rate adaptation has been the most significant function of video transcoding techniques. The idea of compressed video bit rate adaptation is initiated by the applications of transmitting pre-encoded video streams over heterogeneous networks. When connecting two transmission media, the channel capacities of the outgoing channel may be less than those of the incoming channel, so that bit rate adaptation is necessary before sending the video bitstream over heterogeneous channels. In applications such as video on demand, where video is off-line encoded for later transmission, the channel characteristics through which the resulting bitstream will be transmitted might be unknown. Through video transcoding, the bit rate of pre-encoded videos can be dynamically adapted to the obtainable bandwidth and variable communication circumstances. In most bit adaptation cases, a pre-encoded video with high bit rate and fine visual quality will be converted into low bit rate video with elegantly degraded visual quality.

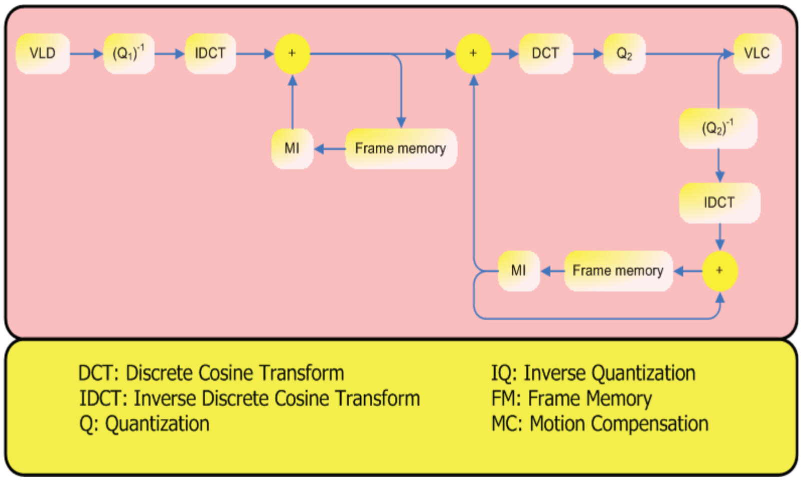
The second function is frame size conversion. Video spatial resolution downscaling is significant since most current handheld devices are characterized by limited screen sizes. By inserting a downscaling filter in the transcoder, the resolution of the incoming video can be reduced. For downscaling the video into lower spatial resolution, motion vectors from the incoming video cannot be reused directly, but have to be resampled and downscaled. Based on the updated motion vectors, predictive residues are recalculated and compressed.

The third function is frame rate conversion. To transcode an arriving compressed video bitstream for a low-bandwidth outgoing channel, such as a wireless network, a high transcoding percentage is often necessary. However, high transcoding ratios may result in intolerable video quality when the arriving bitstream is transcoded with the full frame rate as the arriving bitstream. Frame-rate conversion or frame-dropping is often used as an efficient scheme to assign more bits to the remaining frames, so that acceptable quality can be maintained for each frame. In addition, frame-rate conversion is also needed when an end system can only play video at a lower frame rate due to the processing power limit. Frame rate conversion can be simply accomplished by random frame dropping. For instance, dropping every other frame in a sequential order leads to a half rate reduction in the transcoded sequence. When frames are dropped, motion vectors from the arriving video cannot be directly reused because they are pointed to the immediately previous frame. If the previous frame is dropped in the transcoder, the link between two frames is broken and the end decoder will not be able to reconstruct the picture by these motion vectors. Therefore, the transcoder is in charge for calculating new motion vectors that point to the previous un-dropped frames.

TRANSCODING ARCHITECTURES

Generally speaking, transcoding can be defined as the manipulation or conversion of data into another more desirable format. Depending on the particular strategy that is adopted, the transcoder attempts to satisfy network conditions or user requirements in various ways. In the context of video transmission, compression standards are needed to reduce

Figure 1. Cascaded pixel domain transcoder



the amount of bandwidth that is required by the network. Since the delivery system must accommodate various transmission and load constraints, it is sometimes necessary to further convert the already compressed bitstream before transmission.

The simplest way to develop a video transcoder is by directly cascading a source video decoder with a destination video encoder, which is called the cascaded pixel domain transcoder (Youn, & Sun, 2000). Without using common information, this direct approach needs to fully decode input video and re-encode the decoded video by an encoder with different characteristics as described in Figure 1. Obviously, this direct approach is usually computationally intensive. The architecture is flexible, because the compressed video is first decoded into raw pixels, hence a lot of operations can be performed on the decoded video. However, as we mentioned earlier, the direct implementation of the cascaded pixel domain transcoder is not desirable because it requires high complexity of implementation.

The alternative architecture for transcoding is an open-loop transcoding in which the incoming bit rate is downscaled by modifying the discrete cosine transform (DCT) coefficients. For example, the DCT coefficients can be truncated, requantized, or partially discarded in the optimal sense (Sun, Vetro, Bao, & Poon, 1997; Eleftheriadis & Anastassiou, 1995) to achieve the desirable lower bit rate. In the open-loop transcoding, because the transcoding is carried out in the coded domain where complete decoding and re-encoding are not required, it is possible to construct a simple and fast transcoder. However, open-loop transcoding can produce "drift" degradations due to mismatched reconstructed pictures in the front-encoder and the end-decoder, which often result in an unacceptable video quality.

MOBILE AND WIRELESS VIDEO TRANSCODING SYSTEM ARCHITECTURE

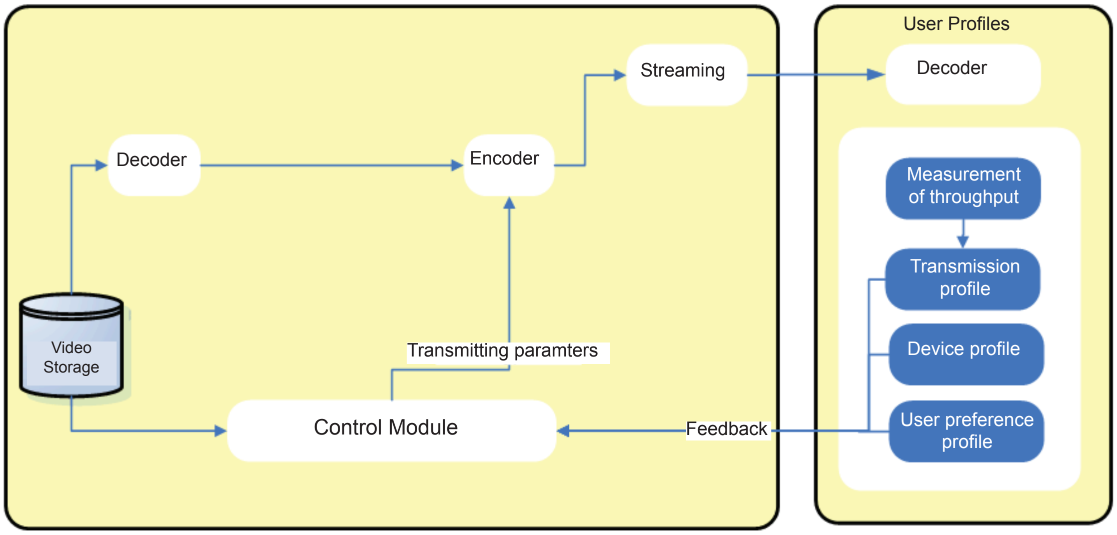
In order to state the mechanism to deploy a transcoding scheme for high-performance transcoding in mobile or wireless communication, we propose a general system architecture for mobile and wireless video transcoding.

In order to enable users to access video through wireless network and handheld devices, we propose using a transcoder as an intermediate node to dynamically convert the video according to user devices and network connections. In this approach, a content provider provides only one video stream, which is pre-encoded at high quality. The video transcoding performs transcoding dynamically for each user and provides converted video streams. The architecture of the video transcoding system is illustrated in Figure 2, as the proposed system contains four important parts: user profile, video transcoder, control module, and pre-encoded video content. In this system, the transcoder is integrated into the video streaming server. It can also be placed as an intermediate node along the transmission path.

User Profile

The user profile maintains several profile modules. The handheld device profile includes hardware and software information on the devices, such as display size, processing power, storage capability, and decoder information. The user preference profile includes user preference information such as user-preferred display size, user-preferred video presentation, and user-preferred video player behavior. The communication profile will measure the download

Figure 2. Overview of mobile and wireless video transcoding system architecture



throughput and update the communication conditions. The information included in the device profile and user profile will be transmitted to the video stream server before the start of the video transmission session. The information included in the transmission profile will be sent to the stream server periodically to control the bit rate adaptation in the transcoder. All information in the user profile is used for parameters in the transcoding process. Table 1 is a clear example for user profile information.

Video Transcoder

The video transcoder is the actual conversion engine of a video stream. It decodes a video stream, which is pre-encoded at high quality and stored in the video source, and then performs transcoding according to our proposed scheme. Our results show that our proposed scheme has very high performance in terms of visual quality. They are comparable to results which can be achieved by full-scale motion estimation-based transcoding. When fast transcoding architectures are used, it is possible to execute transcoding in real time. Thus we can provide the handheld device user a smooth, online video presentation.

Table 1. User profile information

Frame rate
Pixel size
Monitor resolution
Graphic engine capability
CPU usage
Memory usage
Visual objects summary
Visual combination report
Audio object summary
Scene description level

Control Module

The control module is responsible for creating a transcoding scheme according to the user profile and other information. The transcoding scheme will include some transcoding parameters. In order to decide appropriate transcoding parameters, decisions must be made by considering all of the factors adaptively. For example, when connection throughput is low, the bit rate of the video needs to be converted. At the same time, in order to ensure video quality, the frame rate of the video also needs to be reduced. This way, each frame will have enough bit budgets to maintain tolerable visual quality.

CONCLUSION

In this article, several typical video transcoding architectures and major applications of video transcoding have been we reviewed. We identify issues involved in access video streams through handheld devices and wireless networks. We state that the main functions of this transcoding include frame size downscaling, frame rate conversion, bit rate adaptation, color conversion, and so forth. To handle these issues, a few video transcoding system architectures have been proposed; in particular we introduce a system architecture recently suggested by us for intelligently transcoding pre-encoded video for different user devices and network connections in a wireless network environment.

REFERENCES

Correia, P., Faria, S. M., & Assuncao, P. A. (2001). Matching MPEG-1/2 coded video to mobile applications. *Proceedings of the 4th International Symposium on Wireless Personal Multimedia Communications* (Vol. 2, pp. 699-704). Aalborg, Denmark.

Eleftheriadis, A., & Anastassiou, D. (1995). Constrained and general dynamic rate shaping of compressed digital video. *Proceedings of the IEEE International Conference on Image Processing*, Washington, DC.

Eleftheriadis, A., & Anastassiou, D. (1995). Constrained and general dynamic rate shaping of compressed digital video. *Proceedings of the IEEE International Conference on Image Processing,* Washington, DC.

Han, R., Bhagwat, P., LaMaire, R., Mummert, T., Perret, V., & Rubas, J. (1998). Dynamic adaptation in an image transcoding proxy for mobile Web browsing. *IEEE Personal Communications,* 8-17.

Keesman, G. (1996). Transcoding of MPEG bitstreams. *Signal Processing Image Communications, 8,* 481-500.

Mitchell, J., Pennebaker, Fogg, W., Chad, E., & LeGall, J. D. (1996). *MPEG video: Compression standard* (1st ed.). New York: Chapman and Hall.

Shanableh, T., & Ghanbari, M.. (2000). Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats. *IEEE Transactions on Multimedia, 2*(2), 101-110.

Sun, H., Vetro, A., Bao, J., & Poon, T. (1997). A new approach for memory-efficient ATV decoding. *IEEE Transactions on Consumer Electronics, 43,* 517-525.

Warabino, T., Ota, S., Morikawa, D., & Ohashi, M. (2000). Video transcoding proxy for 3G wireless mobile Internet access. *IEEE Communications Magazine,* 66-71.

Youn, J., & Sun, M.-T. (2000). Video transcoding with H.263 bit-streams. *Journal of Visual Communication and Image Representation,* 11.

KEY TERMS

Bit Rate: The transmission speed of binary coded data.

CODEC, Decoder, and Encoder: A device or program capable of performing transformations on a data stream or signal. Codecs can both put the stream or signal into an encoded form (often for transmission, storage, or encryption) and retrieve or decode that form for viewing or manipulation in a format more appropriate for these operations.

Discrete Cosine Transform (DCT): An algorithm that is widely used for data compression. Similar to Fast Fourier Transform, DCT converts data into sets of frequencies. The first frequencies in the set are the most meaningful; the latter, the least. To compress data, the least meaningful frequencies are stripped away based on allowable resolution loss. DCT is used to compress JPEG, MPEG, and H.263 frames.

Mobile Network: The network through which mobile devices communicate. This contains base stations, also known as cell sites, which are in turn linked to the conventional telephone network.

MPEG: The Moving Picture Experts Group is a working group of ISO/IEC charged with the development of video and audio encoding standards.

Multimedia: The use of several different media data elements (text, audio, graphics, animation, video, and interactivity) to express information.

Quality of Service (QoS): The probability of the telecommunication network meeting a given traffic contract, or in many cases, the probability of a packet succeeding in passing between two points in the network.

Mobile Virtual Communities

Christo El Morr
York University, Canada

INTRODUCTION

The adoption of mobile phone technology on a large scale in today's societies turned mobile phones into a universal tool. Phone companies are deploying 3G mobile technology and planning for 4G; nevertheless, the "killer" applications are yet to be developed. Meanwhile, *mobile virtual communities* (MVCs) are emerging, and their applications are diverse: they range from education, to entertainment and lifestyle. Our vision is that mobile virtual communities will be a major trend and could create a momentum for 3G and 4G mobile phone applications. In this article we analyze the different types of mobile virtual communities, and we draw some research perspectives and applications.

BACKGROUND

Computer-supported collaborative work (CSCW) is a multidisciplinary field of research that incorporates people from computing, sociology, psychology, economy, as well as other fields. CSCW research goes back to the 1980s; indeed the first CSCW conference goes back to 1986 (Grudin, 1994). CSCW research aims to study the different human aspects of a group of people working together (a community). Research in CSCW strives to understand how people collaborate together in groups and organizations, and to analyze how computers affect their way of work and how computing can support collaboration between members of a community.

Online communities emerged in beginning of the 1990s with the development of the World Wide Web. Online communities are part of the CSCW field of research, and they can be considered as a kind of a social system. In general, we can identify two types of *social systems* (Preece, 2000): the organization type and the association type. The organization type is designed for a specific aim, and the association type is formed out of individuals' dedication for shared objectives or beliefs. An *online community* is a kind of social system that consists of: (a) socially interacting people; (b) performing special roles or satisfying their needs; (c) a purpose, which is the reason behind the community; (d) policies to govern people interaction; and (e) a computer system that supports social interaction (Weissman, 2000).

Virtual communities (VCs) are online communities where the meeting place is virtual; in fact, the community members in a virtual community are not in the same physical place, rather they are "associating" in a virtual space that is the Internet. *Mobility* emerged in the 1990s and added other aspects to virtual communities. People started to use mobile technology to communicate, to notify each other about events, and to collaborate on common objectives while they are on the move—mobile virtual communities (MVCs) emerged. Mobility shaped new challenges on the technological level and opened new opportunities on the application level; furthermore it incurred new aspects/changes in the community environment that impact the way people interact together and collaborate. The need to understand the types of mobile virtual communities and to explore their application perspectives is of major importance, and a lot of effort is still to be done in this regard. In the next two sections we will first define the different types of virtual communities and then draw some suggestions on pending questions and research perspectives.

TYPES OF MOBILE VIRTUAL COMMUNITIES

To determine the different types of mobile virtual communities, we will begin by looking into the different research areas in the field. Obviously, an overview of MVC research cannot be exhaustive but mostly indicative of the research directions that are taking place. Indeed, MVC research does cover a wide variety of areas such as the infrastructure needed to enable MVC applications, the diverse applications that can serve mobile users/consumers, and the user needs elicitation and user experience analysis.

Some research projects cover the *technology needs* of virtual communities and are concerned with investigating the right hardware and software technologies that could help in establishing a rich and seamless mobile practice in a community. This research area is involved in establishing the infrastructure requirements such as the network and services design, as well as the platform design (Kaji, Ragab, Ono, & Mori, 2002; Pedro Sousa & Garlan, 2002). It also tackles the user interface usability, the wearable devices' impact on the field, and the opportunities that intelligent mobile agents' technology can offer to the MVC experience. *Work*-related research is more oriented towards determining the impact of mobility on the ways people conduct collaboration at work

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

and the kind of influence it has on their workflow (Geisler & Golden, 2003). Other research interest relates to *education* and is concerned with mobility impact on the learning experience. Even though e-learning has been around for a while, mobility added a component to e-learning communities that is pushing learners to further their educational experience, since it enables learners to have access to information while they are on the move (e.g., in the case of field trips) (Farooq, Schafer, Rosson, & Caroll, 2002). In this respect, researchers investigate the way in which collaborative and matchmaking tools can support e-learning. Recently we have witnessed a very promising research orientation in the *entertainment* field. One thinks of mobile gaming, which started to take more and more interest in the last few years; and the huge success of the Apple iPod shows the potential of music virtual communities, while emergence of the interactive TV (iTV) as a community support is opening the way for new opportunities (CHI2006, 2006). Another type of research activity is *lifestyle*-related research that strives to organize and simplify users' daily activities. Visiting a city (tourism), collaborating with colleagues (study, work), and organizing leisure time with friends are all examples of such lifestyle communities (Brown et al., 2005; Silverstone & Sujon, 2005). The most recent aspect of virtual communities is in health care. Indeed, *health* collaborative communities such as the COSMOS project are very recent in MVC research. COSMOS, for example, proposes to build a virtual community to support cancer patients (Leimeister, Daum, & Krcmar, 2004), though mobility is not envisaged yet in the project. Research in the health field is very promising (Johnson & Ambrose, 2006). Finally, *security* as well as trust models are still a universal concern in virtual communities and more specifically in health (Sillence, Briggs, Fishwick, & Harris, 2004).

These different research domains suggest that virtual communities have evolved and are now covering diverse areas such as entertainment and health. While virtual communities emerged from the CSCW field of research, it is becoming a standalone field. The complexity of today's environment at work, the lifestyle that is pushing towards more and diverse entertainment experience, the strive of governments to reduce health care cost, the evolution of the tele-worker concept, as well as the strive of communication companies to generate profit (especially after the stock market crash in 2001)—these are all factors pushing research in different areas of interest. We believe that mobility is a major investigation field for the coming years; indeed, young generations are technology savvy, and mobile technology is widespread driving communication demand. The future picture is one where the mobile environment permits mobile access anytime, anywhere to communities of interest (entertainment, health, education).

RESEARCH PERSPECTIVES

MVCs represent a great potential for mobile network companies; indeed companies are searching for the "killer" application—that is, the application that will embrace the 3G and 4G mobile market; we believe that MVC applications can provide potential services that may drive the demand for the next mobile technologies.

Nevertheless, overcoming the screen size is an issue in mobile devices; this fact is fueling research for a "suitable" graphical user interface and the most appropriate way to organize the letters on the keypad. Mobile usability will continue to stimulate research, but beyond the device the usability of mobile virtual communities is still in its infancy since mobile virtual communities are still emerging.

Several research opportunities in MVCs are still ahead. One can think of the health-related (mobile) virtual communities to see the tremendous potential of MVCs: patient support communities and health promotion communities are only two examples of such potential applications of MVCs; we expect the impact of MVCs in the health field to be tremendous.

The study of members' motivation in order to understand participation and interaction dynamics is still evolving. In this context, artifacts such as the value of the contribution of a member is being proposed to enhance participation (Rashid et al., 2006), and longitudinal studies of newsgroup members' behavior are performed in order to draw conclusions on possible tools to enhance community interaction (Arguello et al., 2006).

Questions related to security and privacy are under investigation; nevertheless we believe that beyond the traditional approach to security and privacy, new research directions related to policymaking (in case of health for example) are still not investigated due to the relatively recent emergence of MVCs and the lack of mobile applications in real life.

We believe that MVC research is an emerging field that needs to be tackled in a multidisciplinary way; notably a joint approach from the CSCW area and the Usability area can lead to several innovative solutions. MVC research is promising to lead to innovative applications if intertwined with upcoming technologies such as 3G/4G mobile phones and pervasive computing.

REFERENCES

Arguello, J., Butler, B., Joyce, L., Kraut, R., Ling, K., & Wang, X. (2006, April 22-27). Talk to me: Foundations for successful individual-group interactions in online communities. *Proceedings of CHI2006* (pp. 959-968), Montreal, Canada.

Brown, B., Chalmers, M., Bell, M., MacColl, I., Hall, M., & Rudman, P. (2005). Sharing the square: Collaborative visiting in the city streets. *Proceedings of the Conference on Human Factors in Computing Systems (CHI2005)*.

CHI2006. (2006). Investigating new user experience challenges in iTV. *Proceedings of the Mobility & Sociability Workshop, Conference for Human-Computer Interaction* (CHI2006). Retrieved from http://soc.kuleuven.be/com/mediac/chi2006workshop/organizers.htm

Farooq, U., Schafer, W., Rosson, M., & Caroll, J. M. (2002). M-education: Bridging the gap of mobile and desktop computing. *Proceedings of the IEEE International Workshop on Wireless and Mobile Technology in Education (WMTE'02)*.

Geisler, C., & Golden, A. (2003). Mobile technologies at the boundary of work and life. *Proceedings of the 2003 Convention of the National Communication Association.*

Grudin, J. (1994). CSCW: History and focus. *IEEE Computer, 27*(5), 19-26.

Johnson, G., & Ambrose, P. (2006). Neo-tribes: The power and potential of online communities in health care. *Communications of the ACM, 49*(1), 107-113.

Kaji, N., Ragab, K., Ono, T., & Mori, K. (2002). Autonomous synchronization technology for achieving real time property in service oriented community system. *Proceedings of the 2nd International Workshop on Autonomous Decentralized Systems.*

Leimeister, J. M., Daum, M., & Krcmar, H. (2004). Towards mobile communities for cancer patients: The case of Krebsgemeinschaft.de. *International Journal of Web-Based Communities, 1*(1).

Pedro Sousa, J., & Garlan, D. (2002). Aura: An architectural framework for user mobility in ubiquitous computing environments. *Software Architecture: System Design, Development, and Maintenance, Proceedings of the 3rd Working IEEE/IFIP Conference on Software Architecture* (WICSA3) (pp. 29-43), Montreal, Canada.

Preece, J. (2000). *Online communities: Designing usability, supporting sociability.* New York: John Wiley & Sons.

Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, P., & Riedl, J. (2006, April 22-27). Motivating participation by displaying the value of contribution. *Proceedings of CHI2006* (pp. 955-958), Montreal, Canada.

Sillence, E., Briggs, P., Fishwick, L., & Harris, P. (2004). *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 663-670), Vienna, Austria.

Silverstone, R., & Sujon, Z. (2005, February). *Urban tapestries: Experimental ethnography, technological identities and place.* An LSE Electronic Working Paper. Retrieved from http://www.lse.ac.uk/collections/media@lse/pdf/EWP7.pdf

Weissman, D. (2000). *A social ontology.* New Haven, CT: Yale University Press.

KEY TERMS

Computer-Supported Collaborative Work (CSCW): A multidisciplinary field that studies the aspects in which computer systems can support people working in collaboration.

Health Promotion: An approach that aims to improve the health of a population by providing information that allows on to make informed decisions about one's own health.

Mobile Virtual Community (MVC): A virtual community where members are on the move.

Online Community: A community where members are communicating using Web-based technology.

Tele-Worker: A worker that is able to perform his or her job tasks remotely (off-site), as in the case of people working at home for a company.

3G/4G: Third-generation and fourth-generation mobile communication technologies that enable access to high bandwidth over mobile phones.

Virtual Community: A community where members are not present in the same physical place.

Mobile-Based Advertising in Japan

Shintaro Okazaki
Autonomous University of Madrid, Spain

INTRODUCTION

The Internet-enabled mobile handset has rapidly achieved worldwide penetration. Combining personal telephony and sophisticated technologies, the mobile Internet has opened new opportunities for offering a diverse range of services, including interactive advertising. In Japan, D2 Communications offers various forms of mobile advertising services delivering promotional information from advertisers.

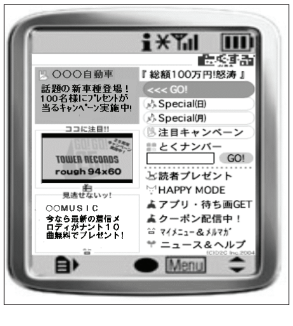
For example, Message F (Free) delivers text-based information to a designated inbox of registered users, who are exempt from the normal packet transmission charges. It can enable highly effective communication due to its ability to target selected demographic segments by region, gender, age, and so forth. In July 2005, D2 Communications began an image attachment service for users of Message F, supporting the transmission of images, logos, and other visual effects, up to 8KB (192 × 192 pixels in JPEG or GIF), as well as text.

This was to be enhanced further in 2006, when Japanese broadcasters began mobile digital broadcasting. The major carrier, NTT DoCoMo, already announced the development of the 3G FOMA(R) P901iTV, which will be "DoCoMo's first mobile handset to receive terrestrial digital broadcasting signals, in addition to conventional analog signals" (NTT DoCoMo, 2005). Therefore, leading mobile advertisers will take advantage of three basic elements in PC-based interactive advertising in mobile devices: static, animated, and broadcast images.

Pull-Type Advertising

In pull-type advertising, messages are displayed to users who voluntarily enter sites (Andersson & Nilsson, 2000) and decide whether to access further information. Consequently, wireless advertisers must improve consumer response and acceptance (Carat Interactive, 2002) because users are unlikely to click banners unless they believe that the content will prove useful, credible, and valuable. Consumers' acceptance, and their perceptions of the delivered content of wireless advertising, are crucial (Carat Interactive, 2002).

Figure 1. Tokusuru menu


○○○自動車
話題の新車種登場！100名様にﾌﾟﾚｾﾞﾝﾄが当るｷｬﾝﾍﾟｰﾝ実施中!
○○MUSIC
今なら最新の着信メロディがナント10曲無料でプレゼント!

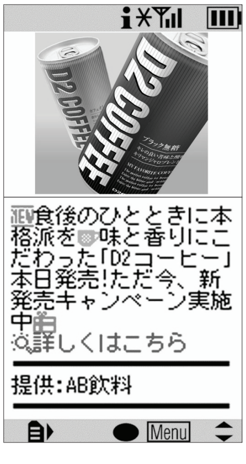
In this vein, D2 Communications offers a pull-type advertising platform called "Tokusuru Menu" (which means "beneficial menu" in Japanese), which provides various text banner ads for promotional campaigns, discount coupons, presents, and so forth. This service requires no registration, and any mobile users can freely access it by selecting No. 4 on the i-menu of an i-mode phone (one of DoCoMo's official sites). Then, they can click and go to the detailed information site. As a result, click-through and call-through rates (almost 15%) are much higher than those of the wired Internet (2-3%) (D2 Communications, 2003; Mizukoshi, Okino, & Tardy, 2001). This popular site attracts an average of 3.5 million people monthly.

Push-Type Advertising

Message F (Free)

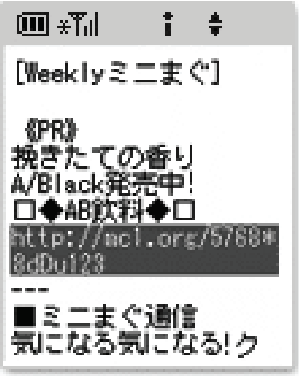
This push advertising delivers selected promotional information from advertisers, exclusively to users who have opted in to receive the service. It is delivered to a designated

Figure 2. Message F


NEW食後のひとときに本格派を味と香りにこだわった「D2コーヒー」本日発売!ただ今、新発売キャンペーン実施中
詳しくはこちら
提供:AB飲料

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 3. Mobile mail advertising

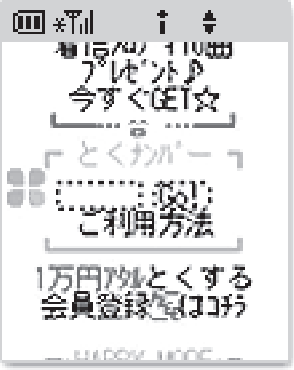


[Weeklyミニまぐ]

《PR》
挽きたての香り
A/Black発売中!
□◆AB飲料◆□
http://mcl.org/5768*
8dDu123

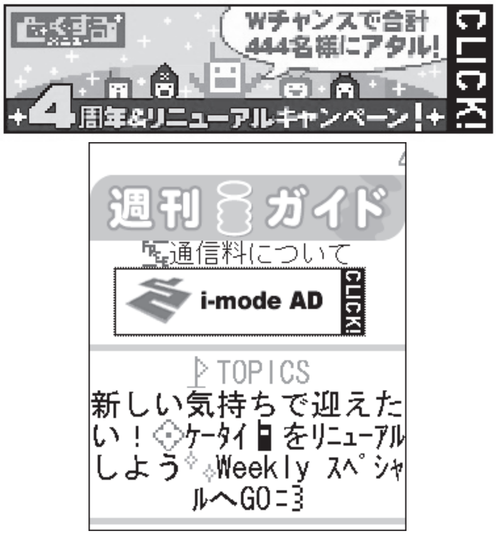
■ミニまぐ通信
気になる気になる!ク

Figure 5. Toku number



プレゼント♪
今すぐGET☆
とくナンバー
ご利用方法
1万円アタルとくする
会員登録はコチラ

Figure 4. Mobile picture advertising



通信料について

TOPICS
新しい気持ちで迎えたい!ケータイをリニューアルしよう Weekly スペシャルへGO=3

"Message F" inbox, and users are exempt from the normal packet transmission charges. It can enable highly effective communication, due to its ability to target selected demographic segments by region, gender, age, and so on.

Mobile Mail Advertising

This advertising is delivered in text format and inserted into mail-magazines' headers. The target customers are "opt-in" users who have registered their regular subscription. User responses often follow directly upon delivery. This category of advertising elicits a high level of response.

Banner Advertising

In this form of advertising, advertisers can send a static image as banner to the sites linked with i-menu. This banner ad occupies approximately one-quarter of a micro-browser screen. Usually it provides a good exposure of logos and other visual elements, and can thus be used as an attention-grabbing tool. According to the most recent statistics, approximately 1.8 billion advertising banners have been sent so far, with an average click-through rate of 4.6%.

Toku Number

This advertising has been developed as one of the direct marketing strategies via mobile devices. Basically, this is a short code (or "number," as reflected in its name) that enables users to connect directly to a designated mobile campaign site or clients' mobile site. In Japan, each mobile operator carries an informative site, such as *Tokusuru Menu* for i-mode, *Tokusuru Info* of EzWeb, or *Tokusuru Information Board* of Vodafone Live, but users can insert this code in any of these Internet sites, because it is a standardized form or common campaign code. This code is frequently used in a cross-media campaigns: a given campaign's Toku number appears in television or print media so that mobile users can input the code to their mobile handsets directly.

Electronic Wallet

In July 2004, NTT DoCoMo's new i-mode service "FeliCa" became available. FeliCa is a multi-functional electronic wallet with contactless electronic IC chips developed by Sony. In combination with NTT DoCoMo's "i-appli" (Java-based applications), users can use FeliCa for diverse transactions, such as commuter pass, electronic money, membership card, and movie tickets, among others, simply by waving their phone in front of enabled sensors (IT Media Mobile, 2003). Figure 2 shows some micro-browser screens of FeliCa. FeliCa combines a wireless Internet service with electronic financial transactions, and offers three principal functions in its main menu: e-Wallet, e-Card, and e-Ticket.

This breakthrough technology is expected to expand greatly the uses and productivity of wireless advertising platforms by enabling users not only to access promotional benefits in banner ads, but also to make use of them. For example, when advertisers display discount coupons in a wireless advertising platform, users can browse the information, transfer it to FeliCa with one click, and then redeem the coupon at a point of purchase using FeliCa. At convenience stores, payments can be made using electronic money, "Edy" (developed by Bit Wallet, Inc.), simply by waving the handset at the cash register (IT Media Mobile, 2004a). In July 2004, "Toho Cinemas," a movie theater chain, started a wireless ticketing service through FeliCa-enabled mobile handsets (IT Media Mobile, 2004b), while NTT DoCoMo and two international payment companies, JCB and AEON Credit Service, announced that they are jointly to develop a quick payment solution for FeliCa-enabled handsets (NTT DoCoMo, 2004).

Furthermore, FeliCa technology can also enhance the feasibility of applying the global positioning system (GPS) to the wireless Internet service. For example, on an extended menu of i-mode, "i-area" includes a diverse range of location-based services: weather news, restaurant guide, local hotel information, zoomable maps with an address-finder function, and traffic updates and estimation of travel times (Sadeh, 2002). In the future, advertisers will be able to beam "real-time offers" to subscribers in targeted, location-specific campaigns, with benefits redeemable in a quick and timely manner.

CONCLUSION

This article summarizes the current practices of mobile advertising in Japan. These services are mainly offered through i-mode of NTT DoCoMo, which has been "exported" to European countries, as well as the United States. Although the level of acceptance of mobile advertising remains limited, the use of mobile advertising has become increasingly popular among consumer goods manufacturers. At the same time, the information provided in this article will be a useful reference for international advertisers and marketers, given that the mobile Internet has been achieving a high penetration worldwide. In Japan, major mobile advertising agencies started an image attachment service to their messaging services in 2005, supporting the transmission of images, logos, and other visual effects. In 2006, mobile carriers will begin to launch a mobile handset which can receive terrestrial digital broadcasting signals, in addition to conventional analogue signals, enabling firms to take advantage of three basic elements in PC-based interactive advertising in mobile devices: static, animated, and broadcast images. This terrestrial digital broadcasting is expected to be a breakthrough change in terms of the advancement of mobile-based advertising strategies in Japan.

REFERENCES

Carat Interactive. (2002). *The future of wireless marketing.* Retrieved from http://www.caratinteractive.com/

D2 Communication. (2002). *Our lines of business.* Retrieved from http://www.d2c.co.jp/english/business_e/ main.html

IT Media Mobile. (2003). FeliCa tosai de keitai wa kokawaru (How FeliCa changes the use of mobile handsets). *Mobile: News,* (October 27). Retrieved from http://www.itmedia.co.jp/mobile/0310/27/n_doso2.html (in Japanese).

IT Media Mobile. (2004a). Mieta, FeliCa keitai no hon service (Understanding main services of FeliCa-enabled handsets). *Mobile: News,* (June 16). Retrieved from http://www.itmedia.co.jp/mobile/articles/0406/16/news101.html (in Japanese).

IT Media Mobile. (2004b). FeliCa de eiga-ticket hakken wo 7 gatsu 10 ka kaishi (Starting to sell move tickets via FeliCa from 10th July. *Mobile: News,* (July 9). Retrieved from http://www.itmedia.co.jp/mobile/articles/0407/09/news056.html (in Japanese).

Mizukoshi, Y., Okino, K., & Tardy, O. (2001). Lessons from Japan. *Telephony,* (January 15), 92-96.

NTT DoCoMo. (2003). *Introducing i-mode.* Retrieved from http://www.nttdocomo.com/home.html

NTT DoCoMo. (2004a, July 9). *i-mode users outside Japan exceed 3 million.* Retrieved from http://www.nttdocomo.com/article/?no=Mzc4LzEzMTM=

NTT DoCoMo. (2004b, July 20). *JCB and AEON develop new contactless payment solution, 'QUICPay', for cards and NTT DoCoMo's mobile phones compatible with i-mode FeliCa service.* Retrieved from http://www.nttdocomo.com/article/?no=Mzc4LzEzNTk=

Sadeh, N. (2002). *M-commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

KEY TERMS

Electronic Wallet: In July 2004, NTT DoCoMo's new i-mode service "FeliCa" became available. FeliCa is a multi-functional electronic wallet with contactless electronic IC chips developed by Sony. In combination with NTT DoCoMo's "i-appli" (Java-based applications), users can use FeliCa for diverse transactions, such as commuter pass, electronic money, membership card, and movie tickets, among others, simply by waving their phone in front of enabled sensors (IT Media Mobile, 2003).

Message F (Free): A push-type advertising offered in i-mode which delivers promotional information exclusively to users who have opted in to receive the service. It is delivered to a designated "Message F" inbox, and users are exempted from the normal packet transmission charges.

Mobile Banner Advertising: A static or animated image can be used as a banner to the sites linked to the i-menu. Generally, it provides a good exposure of logos and other visual elements, and thus can be used as an attention-grabbing tool.

Tokusuru Info: An equivalent service to Tokusuru Menu in EZWeb.

Tokusuru Information Board: An equivalent service to Tokusuru Menu in Vodafone Live.

Tokusuru Menu: A pull-type advertising platform offered by D2 Communications. This is one of the "official" i-mode sites, which provides various text banner ads for promotional campaigns, discount coupons, presents, and so forth. This service requires no registration, and any mobile users can freely access if by selecting No. 4 on the i-menu of an i-mode phone. "Tokusuru" means "beneficial" in Japanese.

Mobile-Based Research Methods

Shintaro Okazaki
Autonomous University of Madrid, Spain

Akihisa Katsukura
Dentsu Inc., Japan

Mamoru Nishiyama
Dentsu Communication Institute Inc., Japan

INTRODUCTION

The Internet-enabled mobile handset has rapidly achieved worldwide penetration. Combining personal telephony and sophisticated technologies, the mobile Internet has opened new opportunities for offering a diverse range of services, including online surveys. In particular, the mobile-based messaging service can be used as a practical tool for transmitting questionnaires and collecting responses (see Figure 1). This method also offers a solution to researchers who have begun to recognize that an important question remains unanswered in mobile research: How can we actually "capture" mobile Internet users? So far, the majority of empirical studies in this area have used the traditional pen-and-pencil survey method, while little care has been taken to ensure that the respondents are actual mobile Internet adopters who are capable of evaluating such a new medium. Because of the ubiquitous nature of the mobile device, a conventional questionnaire may be inappropriate for capturing "true" targets, and unlike PC-based e-commerce research, there are several factors to be considered in terms of survey planning and executions.

The aim of this article is to propose a framework of mobile-based survey methodology. Specifically, we attempt to establish guidelines for a questionnaire survey via the mobile device, in terms of cost, questionnaire format, incentives, target respondents, response rate, and data quality.

CRITICAL ISSUES TO BE CONSIDERED

Cost

Prior research indicates that the total cost of Internet-based surveys requires a higher set-up cost in designing, programming, and hosting sites, in comparison with paper-based surveys. However, Internet-based surveys demand neither paper nor postage, thus reducing overall costs by 30-60% (Hanna, Weinberg, Dant, & Berger, 2005). Furthermore, typical online surveys require no expenses related to photocopying, clerical support, and data entry, since the responses can be input into the data analysis software automatically (Llieva, Baron, & Healey, 2002).

While similar benefits can be expected from mobile-based surveys, the wireless device requires a higher connection cost in data collection. This is particularly an issue, because although mobile penetration has been growing worldwide, Internet connection capacity still varies between countries in terms of both infrastructure and handset functionality. For example, SMS seems comparable to the Japanese e-mail service in terms of cross-carrier compatible, text-based, and immediate send-and-deliver, but the simple SMTP-based mail used by DoCoMo, KDDI, and Vodafone offers more practical advantages. With regard to pricing, Scuka (2003) notes:

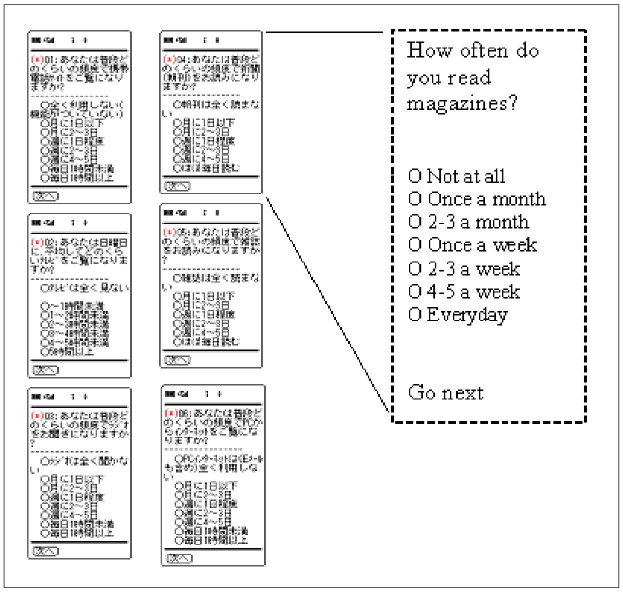
> *It costs 0.3 yen to send 1 packet on DoCoMo's 2.5G i-mode network in Japan; 1 packet is 128 bytes, so a typical simple e-mail might only cost 1 yen (and even cheaper on 3G, where the average packet price is around 0.06 yen/packet). SMS messages, in contrast, cost 19 Euro cents to send 160 characters on all four carriers in Germany (T-Mobile, Vodafone, E-Plus, and O2).*

Response Format

The choice of response format is another important question. Here, two basic issues must be considered: (1) dichotomous vs. multichotomous scale formats, and (2) the wording of the response scale points (Netemeyer, Bearden, & Sharma, 2004). Dichotomous scales have practical advantages, in that they take less time for a respondent to fill in and, therefore, allow more items to be responded to in a short time. However, they have been criticized for their tendency to have highly unbalanced response distributions—that is, all individuals always answering "true" or all individuals always answering "false" (Comrey, 1988). Furthermore, any one item for a dichotomous scale produces only limited covariance with any other item because of the binary format, and overall scale variance will be very limited. Multichotomous scales overcome these shortcomings, in that they create more

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Example of mobile-based online questionnaire (Source: D2 Communications). An English translation is shown on the right side.



How often do you read magazines?

○ Not at all
○ Once a month
○ 2-3 a month
○ Once a week
○ 2-3 a week
○ 4-5 a week
○ Everyday

Go next

scale variance relative to a dichotomous scale with a similar number of items.

For these reasons, academic researchers tend to prefer multichotomous formats, while practitioners are more likely to adopt dichotomous formats. For example, our depth-interview with an agency practitioner reveals that the reason for the exclusive use of dichotomous response format was its ease of use and minimum response time, which agencies believe are closely related to data quality (Fukada, 2005).

Type of Incentives

In general, it is agreed that the use of monetary incentives has positive effects in increasing the response rate, in both online and off-line surveys (Church, 1993; Yammarino, Skinner, & Childers, 1991). However, what "type" of incentives could increase the response rate seems unclear. For example, Ray, Griggs, and Tabor (2001) found that as much as 57% of the respondents agreed to an exchange for a draw/raffle inclusion, while Comley (2000) reports that such an incentive has little impact in e-mail surveys. A recent experimental study shows that a shorter questionnaire with small lotteries with higher winning chances would produce a higher response rate, but a longer questionnaire also could generate a reasonable response rate if vouchers were promised (Deutskens et al., 2004).

In the case of a mobile-based survey, a barcode coupon has been widely used in various cases of push marketing (Senden Kaigi, 2004). Furthermore, a free download of screen images, ring-tones, online games, product samples, electronic coupons, and a sweepstakes competition have been tested as effective incentives to encourage participation (Senden Kaigi, 2004). In particular, the free content download (in particular, ring-tone and screen image) and sweepstakes competition offer practical and economical solutions, given that they represent two of the most popular mobile Internet usages in many markets (Dano, 2002; Kim et al., 2004; Harris, Rettie, & Kwan, 2005).

Target Respondents

One of the major problems in m-commerce research is that researchers are often unable to ensure whether target consumers actually have sufficient experience and are capable of providing reliable responses. Empirical studies that used "general" consumer samples seldom conditioned their experience in accessing the mobile Internet, and did not ensure that the target respondents had actually adopted the mobile device for Internet connection. Also, we need to cover a wider range of consumer groups, in both demographic and socioeconomic characteristics. In particular, in an m-commerce context, it is extremely important that the survey examine those who regularly access the mobile Internet, read promotional messages, and even access the linked sites to obtain further information, on a daily basis (Okazaki, 2005). This clearly relates to the respondents' perceived ability to use the mobile Internet, because mobile-based questionnaires can only be returned from those who are willing to reply via mobile device. This is significantly different from PC-based online surveys, where researchers can offer alternative response options such as fax or postal mail (Truell, 2003).

Response Rate

The response rate of surveys has often been used to assess data quality (Shermis & Lombard, 1999). According to Comley (2000), the overall response rate of all virtual surveys in 1999 ranged from 15% to 29%. However, considerable discrepancies have been found between similar studies in similar periods (Ray et al., 2001; Virtual Surveys, 2001; Wygant & Lindorf, 1999). In many cases, the response rate was found to depend upon the researched topic, which may or may not encourage active participation (Sheehan & McMillan, 1999; Ray et al., 2001).

These figures have an important implication for the computation method of response rate, since a mobile-based questionnaire is normally sent via a push messaging service to the "opt-in" users. Therefore, the target base is a total population of a given area, and whether recipients actually click and open the questionnaire is crucial. At the same

Table 1. Principal component analysis results

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Creativity							
Eye-catching	.670						
Esthetic	.632						
Original	.583						
Easy to understand	.452						
Entertainment							
Fun		.634					
Pleasure		.626					
Convenience							
Simple			.744				
Cool			.657				
Easy to use			.596				
Speed							
Time-saving				.681			
Advanced				.669			
Quick				.485			
Pragmatism							
Advantageous					.792		
Appealing					.792		
Curiosity							
Interesting						.767	
Useful						.465	
Trust							
Trustworthy							.745
Well-crafted							.743

time, empirical evidence shows that the click-through rate of push-type messages has been quite stable (D2 Communications, 2005). A question then arises as to the base on which we should compute the response rate. Should the response rate be calculated on the basis of the total population or on those who actually click the message? The former method seems inappropriate, because we would include in the calculation of the response rate those who are unlikely to open the questionnaire.

PILOT STUDY

A pilot study was carried out in July 2005 in an attempt to examine the respondents' perceptions of mobile-based advertising campaigns. Forty thousand mobile users were randomly chosen from the agency's customer database, to which a structured questionnaire was sent directly via mobile device. As an incentive, a book coupon was offered. The questionnaire consisted mainly of 23 adjectives that represent consumers' psychological motivations to willingly receive, open, and read mobile-based advertising campaigns. In addition, media habits as well as demographic information were collected. In all question items, the dichotomous response format was used, and consumers were asked to check the appropriate box to indicate whether an adjective appropriately expressed his or her motivation to accept mobile-based advertising campaigns.

As a result, 1,401 people responded to the survey. The results were converted into fictitious variables, by assigning 1 to "checked" (which means "yes") and 0 to "unchecked" (which means "no"). A principal component analysis with Varimax rotation was then performed. The results are shown in Table 1. The analysis produced a clear-cut seven-factor solution, with 50% of the total variance explained. Creativity, entertainment, and convenience were the three

Table 2. Summary of two-step cluster analysis

Cluster	Characteristics
Cluster 1	This cluster mainly consists of male white-collar workers between the 20s and early 40s. Their occupations include administrative, research, and professional workers.
Cluster 2	This cluster consists of students who are under 25 years old. The proportions of female and male respondents are almost equal.
Cluster 3	This cluster mainly consists of housewives between early 20s and early 30s, including some part-time workers. This cluster is least willing to accept mobile-based advertising campaigns.
Cluster 4	Three-quarters of this cluster consist of "working women" who are between early 20s and early 30s. More than half of this cluster has occupations such as sales and service (30%), administrative, clerical, and office jobs. This cluster is most willing to accept mobile-based advertising campaigns.

most accentuated factors, which seems to corroborate prior e-commerce research on users and gratification theory (Lin, 1999). Speed, pragmatism, curiosity, and trust were the remaining factors.

Next, to identify possible segments of mobile Internet users, we performed two-step cluster analysis via SPSS 13.0. We included respondents' media habits and demographic variables in the analysis. As a result, we extracted four clusters, the major characteristics of which are summarized in Table 2. Clearly, gender is a key factor in differentiating mobile Internet behavior, in terms of the willingness to accept mobile-based advertising campaigns.

IMPLICATIONS

E-mail or Web-based questionnaire surveys have been widely used by marketing researchers because of their low costs and high response quality. Similar benefits can be expected in mobile-based surveys. Although important differences exist across markets in terms of mobile Internet technology and penetration, a mobile messaging service can be effectively used as a carrier of structured questionnaires. Researchers should take into consideration various factors influencing the response rate and data quality, such as cost, choice of response format, type of incentives, and so forth. In particular, future research via mobile device should consider the inclusion of multichotomous response format, so that researchers can apply sophisticated multivariate analysis.

REFERENCES

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing, 16*(1), 14-24.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rate: A meta-analysis. *Public Opinion Quarterly, 57*(1), 62-79.

Comley, P. (2000, April). Pop-up surveys: What works, what doesn't work and what will work in the future. *Proceedings of the ESOMAR Net Effects Internet Conference*, Dublin. Retrieved from http://www.virtualsurveys.com/papers/popup-paper.htm

D2 Communications. (2005, November). Information from depth-interview with the President A. Fujita.

Hanna, R. C., Weinberg, B., Dant, R. P., & Berger, P. D. (2005). Do Internet-based surveys increase personal self-disclosure? *Database Marketing & customer Strategy Management, 12*(4), 342-356.

Harris, P., Rettie, R., & Kwan, C. C. (2005). Adoption and usage of m-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom. *Journal of Electronic Commerce Research, 6*(3), 210-224.

Llieva, J., Baron, S., & Healey, N. M. (2002). Online surveys in marketing research: Pros and cons. *International Journal of Market Research, 44*(3), 361-376.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* Thousand Oaks, CA: Sage.

Okazaki, S. (2006). What do we know about mobile Internet adopters? A cluster analysis. *Information & Management, 43*(2), 127-141.

Ray, N., Griggs, K., & Tabor, S. (2001, April). *Proceedings of the Web-Based Survey Research Workshop (WDSI)*. Retrieved from http://telecomm.boisestate.edu/research/

Scuka, D. (2003). *How Europe really differs from Japan.* Retrieved October 2005 from http://www.mobiliser.org/article?id=68

Sendenkaigi (2004). Mobile marketing & solution. Tokyo: Sendenkaigi.

Sheehan, K. B., & McMillan, S. J. (1999). Response variation in email surveys: An exploration. *Journal of Advertising Research, 39*(4), 45.

Truell, A. D. (2003). Use of Internet tools for survey research. *International Technology, Learning, and Performance Journal, 21*(1), 31-37.

Virtual Surveys. (2001). *E-mail surveys in virtual surveys: Web site research experts.* Retrieved from http://www.virtualsurveys.com/services/email_web.htm

Wygant, S., & Lindorf, R. (1999). Surveying collegiate Net surfers—Web methodology or mythology? *Quirk's Marketing Research Review,* (July). Retrieved from www.quirks.com

Yammarino, F.J., Skinner, S.J., & Childers, T.L. (1991). Understanding mail survey response behavior. *Public Opinion Quarterly, 55*(4), 613-639.

KEY TERMS

Barcode Mobile Coupon: Mobile barcoding can be used in the form of a picture SMS which is delivered to a mobile phone. Recipients save the image, arrive at the destination, and present their barcode SMS to be scanned.

i-mode: A broad range of Internet services for a monthly fee of approximately 3 euro, including e-mail, transaction services (e.g., banking, trading, shopping, ticket reservations, etc.), infotainment services (e.g., news, weather, sports, games, music download, karaoke, etc.), and directory services (e.g., telephone directory, restaurant guide, city information, etc.), which offers more than 3,000 official sites accessible through the i-mode menu.

Opt-In: Process of actively soliciting a permission from a user to which an advertiser sends a promotional message or to collect personal information for marketing purposes.

Opt-Out: Process of giving an explicit notification that a user does not wish to receive any promotional messages or does not wish to have his or her personal information collected for marketing purposes.

Push Messaging Service: Various forms of messaging services are generally offered in the mobile Internet. For example, SMS and WAP push messaging generally allow users to send 100-160 characters, while mobile e-mail in Japanese i-mode allows up to 1,000 characters.

Short Message Service (SMS): A service for sending messages of up to 160 characters to mobile devices.

Mobility and Multimodal User Interfaces

Christopher J. Pavlovski
IBM Corporation, Australia

Stella Mitchell
IBM T. J. Watson Research, USA

INTRODUCTION

Traditional user interface design generally deals with the problem of enhancing the usability of a particular mode of user interaction, and a large body of literature exists concerning the design and implementation of graphical user interfaces. When considering the additional constraints that smaller mobile devices introduce, such as mobile phones and PDAs, an intuitive and heuristic user interface design is more difficult to achieve.

Multimodal user interfaces employ several modes of interaction; this may include text, speech, visual gesture recognition, and haptics. To date, systems that employ speech and text for application interaction appear to be the mainstream multimodal solutions. There is some work on the design of multimodal user interfaces for general mobility accommodating laptops or desktop computers (Sinha & Landay, 2002). However, advances in multimodal technology to accommodate the needs of smaller mobile devices, such as mobile phones and portable digital assistants, are still emerging.

Mobile phones are now commonly equipped with the mechanics for visual browsing of Internet applications, although their small screens and cumbersome text input methods pose usability challenges. The use of a voice interface together with a graphical interface is a natural solution to several challenges that mobile devices present. Such interfaces enable the user to exploit the strengths of each mode in order to make it easier to enter and access data on small devices. Furthermore, the flexibility offered by multiple modes for one application allows users to adapt their interactions based on preference and on environmental setting. For instance, hands-free speech operation may be conducted while driving, whereas graphical interactions can be adopted in noisy surroundings or when private data entry, such as a password, is required in a public environment.

In this article we discuss multimodal technologies that address the technical and usability constraints of the mobile phone or PDA. These environments pose several additional challenges over general mobility solutions. This includes computational strength of the device, bandwidth constraints, and screen size restrictions. We outline the requirements of mobile multimodal solutions involving cellular phones. Drawing upon several trial deployments, we summarize the key designs points from both a technology and usability standpoint, and identify the outstanding problems in these designs. We also outline several future trends in how this technology is being deployed in various application scenarios, ranging from simple voice-activated search engines through to comprehensive mobile office applications.

BACKGROUND

Multimodal interaction is defined as the ability to interact with an application using multiple sensory channels (i.e., tactile, auditory, visual, etc.). For example, a user could provide input by speaking, typing on a keypad, or handwriting, and receive the subsequent response in the form of an audio prompt and/or a visual display. Useful multimodal applications can cover a broad spectrum including tightly synchronized, loosely synchronized, and complementary modes of operation. Synchronization behavior must be defined both for input (the way in which input from separate modes is combined) and for output (the way in which input from one mode is reflected in the output modes). The W3C distinguishes several types of multimodal synchronization for input as follows (W3C, 2003a):

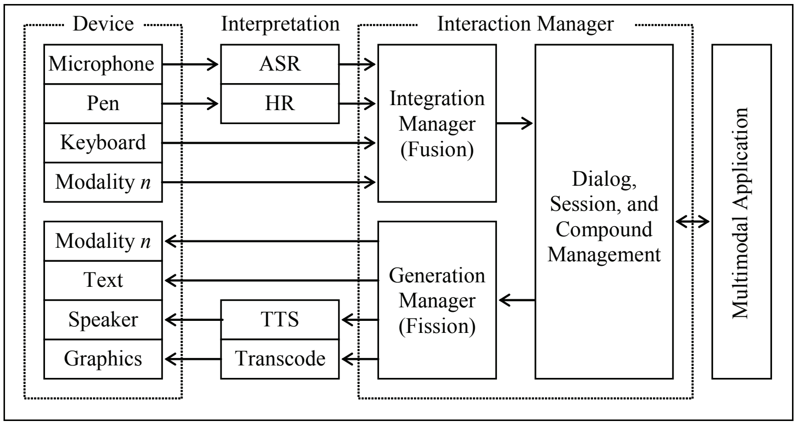
- **Sequential:** Two or more input modalities are available, but only a single modality is available at any given time.
- **Simultaneous:** Allows input from more than one modality at the same time, but each input is acted upon separately in isolation from the others.
- **Composite:** Provides for the integration of input from different modes into one single request.

A general framework for multimodal systems is depicted in Figure 1. This diagram elaborates further on several fundamentals positioned by W3C.

The interaction manager is responsible for combining multiple requests, dialog management, and synchronization. The function of receiving and combining multiple inbound

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Multimodal framework



requests is the responsibility of the *integration manager* subcomponent. Conversely, the generation manager is responsible for distributing multimodal output to all of the respective output channels (modes) via an interpretation layer, which may involve text to speech (TTS) conversion or transcoding of graphical content to accommodate the needs of the target modality. Earlier work in multimodal systems referred to the integration tasks relating to composition and decomposition of requests as *fusion* and *fission* respectively (Coutaz, Nigay, & Salber, 1993).

Speech-based telephone interfaces currently available in the commercial market commonly use varying levels of directed dialog. Directed dialog, as the name implies, employs a style of system prompts that helps to "direct" the user in what to say next. Users are often presented with spoken menu options from which they can make a selection, thus navigating in a controlled manner until the task is completed. Much of the naturalness and power of speech is undermined when the application relies too heavily on the use of directed dialogs. A Natural Language speech interface, which allows the user to phrase their request in a wide variety of ways, reduces the cognitive load since there are no commands to memorize or hierarchies to navigate. A mixed-initiative interface allows the user to share control over the direction of the dialog, making the interaction more efficient for the user.

Device manufacturers can install specialized software or firmware on handsets to enable distributed speech recognition (DSR). DSR technology digitizes the speech signal and sends it over an error-protected data channel to a speech recognizer on a server. Thus the processing is distributed between a terminal client and a server. The accuracy of speech recognition can be better when using DSR because the degradations associated with sending speech over the mobile network, such as low bit rate speech coding and channel transmission errors, are avoided. In addition, DSR allows for the implementation of a multimodal, speech and data, application on devices which do not support simultaneous voice and data connections.

MULTIMODAL TECHNOLOGY IN MOBILITY SYSTEMS

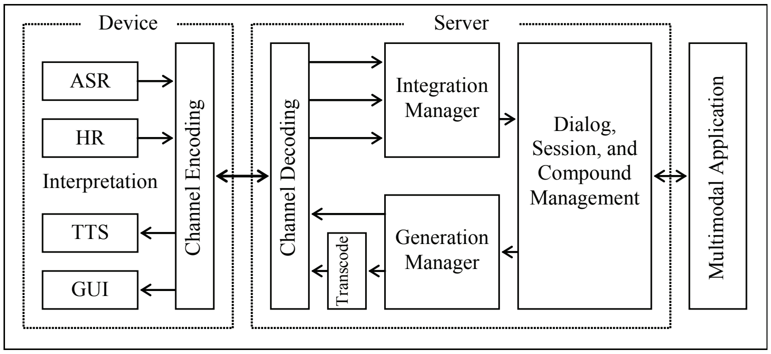
Multimodal applications for small mobile devices must overcome several technical challenges; these include device capability and computational strength, functional and bandwidth constraints of the network, and limitations of the user interface. Present work in multimodality is focused upon the support of two modes of interaction, most typically data (graphics or text) and speech (Kondratova, 2004; Pavlovski, Lai, & Mitchell, 2004a; Hastie, Johnston, & Ehlen, 2002; Kvale, Warakagoda, & Knudsen, 2003). Future trends support additional modes that include visual gesture, lip reading, and haptic responses. In this section we present advances in multimodality supporting speech and data entry, outlining the current state of the technology used and the outstanding challenges yet to be addressed.

Multimodal Architectures

Due to the need to support two or more modes of input and output, the solutions to support multimodal systems are more complex than unimodal systems. Additional capabilities are required to support composite input and output requests, manage multiple application states, and perform session management between devices and the multimodal application services.

There are fundamentally two architectural approaches to constructing multimodal solutions for mobile devices such as mobile phones and PDAs. The most widely investigated architecture appears to involve deployment of an application

Figure 2. Distributed client multimodal architecture



onto the mobile device, using distributed speech recognition (DSR) and a GUI client application. Such a solution may be categorized as *distributed client architecture* (i.e., thick client architecture). The alternative employs the *browser-based architecture* (thin client), where a standard Web browser is the only client software used.

Several proposed solutions have been studied that support distributed client architectures (Kvale, Narada, & Warakagoda, 2005; Klante, Krösche, & Boll, 2004). Figure 2 provides an overview of the key sub-systems and components of the distributed client architecture. The device is required to host several components of the solution, generally including a multimodal client application and distributed multimodal processing technologies. This may include automated speech recognition (ASR), text to speech (TTS), and handwriting recognition (HR); however, typical distributed client designs to date have such components on the server. Communication between the client and server occurs over a single channel; hence the multimodal input is marshaled (encoded) into one communications stream.

There are several advantages of this type of architecture including true synchronicity in delivery of the multimodal output, distributed processing, and richer user interface. Since the multimodal output is combined before transmission to the client, the client application is able to ensure media synchronization—that is, synchronized presentation of content to the user; this is a key problem to address in multimodal systems. The presence of a client application also provides a richer graphical user interface and consistency in user interface output. Additionally, the capability to distribute some processing tasks to the client provides greater flexibility for multimodal applications, particularly in relation to response time. For instance, a multimodal short message service is able to perform speech recognition locally within the device, rather than returning to the server, and display the utterance to the user before sending as a short text message.

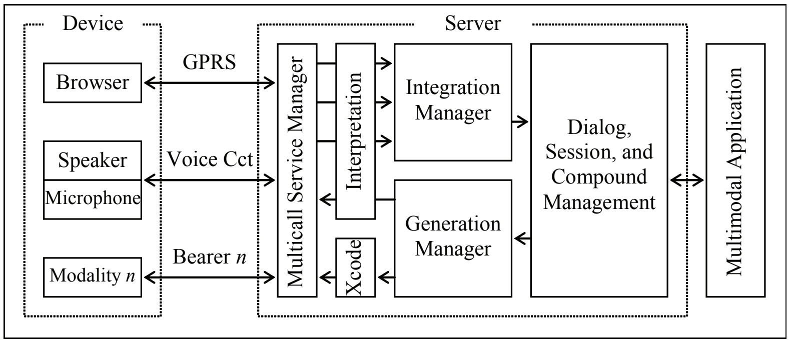
A key drawback to distributed client architectures is the need to manage the client software to be deployed to the mobile device. This is appears to be of particular concern to mobile operators for several reasons. Given the significant numbers of mobile phones in use, software distribution and management to the device is costly. The majority of mobile phone plans are 12 or 18 months, hence new devices are acquired rapidly and continual support becomes cost prohibitive. Additionally, it may not be practical to support all devices available on the market.

Further research suggests that while a limited graphical user interface is of concern to users, in fact the major contributor to improved usability of multimodal user interfaces is the speech interface (Hastie et al., 2002; Pavlovski et al., 2004a). There is also much work on the notion of speech-centric multimodal user interfaces (Johnsen & Kvale, 2005). These observations also support the notion that alternative architectures require consideration.

An alternative approach is a class of multimodal system that does not require client applications to be deployed to the mobile device. Such a framework is also termed a browser-based, or thin client, architecture—only requiring a standard browser on the mobile phone. In order to support the multiple channels, typically speech and data, use is made of a '*Class A*' mobile phone and the multi-call supplementary service, which is only available in 3G networks. Specifically, each device establishes several bearer channels with the server, one per mode of user interaction. For example, simultaneously a voice circuit is established for speech and a data bearer service for the browsing graphical and text content (see Figure 3).

Thin client architectures have been studied for mobile devices. An initial set of building blocks to support browser architectures for sequential multimodal processing was outlined in Niklfeld, Finan, and Pucher (2001). The approach in Chou, Shan, and Li (2003) and Li, Wong, and Guo (2004) supports a thin client architecture, where use is made of the multi-call supplementary service for simultaneous connections between the mobile phone and the multimodal application server; however, only sequential multimodal input and output is supported. A further solution solves the

Figure 3. Browser multimodal architecture



problem of supporting composite multimodal output response (Pavlovski, Wood, Mitchell, & Jones, 2004b), however only sequential simultaneous multimodal input is supported. The MONA project provides a similar capability for thin client, and extends the device range to support both mobile phones and PDA devices using a multimodal presentation server (Anegg, Dangl, & Jank, 2004).

The browser-based architecture would naturally seem more attractive to mobile phone operators given the dynamic and rapid turnover of mobile phones, hence eliminating the need to manage software distribution and complexity attributed to supporting numerous mobile devices. The disadvantage of course is that coordinating additional channels increases the complexity and introduces the possibly of latency or synchronization errors. One paper points out that media synchronization is an impact to multimodal applications (Pavlovski et al., 2004b). This may be accommodated by delaying one of the responses so that both arrive simultaneously at the client device. However, complex multimodal applications are yet to be studied to comprehensively assess the impact of synchronization.

Several key challenges remain for browser-based multimodal architectures. This includes management of channel latency and media synchronicity. A full implementation that supports both input and output multimodal composite requests remains as an outstanding problem. And, support for three or more multimodal channels is yet to be explored. These problems may be solved by using the distributed client architecture and have been demonstrated (Johnsen & Kvale, 2005)—however, at the cost of introducing additional software management overheads and restrictions on device supportability.

Standards and Technology

The W3C organization has defined several multimodal standards, including interaction requirements and an interaction framework (W3C, 2003b). We briefly summarize some further technologies that are specific to multimodal solutions, outlining their applicability to mobile devices.

Extensible Multimodal Annotation (EMMA)

EMMA is a set of XML specifications that represents multimodal user input including speech, pen, and keystroke. The XML is intended to be generated automatically, hence is suitable for use between recognition engines and the interaction manager.

Multimodal Presentation Markup Language (MPML)

The MPML Language is designed for presenting multimodal output on browsers supporting XML. The limited processing capability of phones is accommodated, with a mobile edition defined for J2ME applications (Saeyor, Mukherjee, Uchiyama, & Ishizuka, 2003). As such, the architecture resembles a distributed client solution.

Speech Application Language Tags (SALT)

SALT is a set of extensions to HTML (or XHTML and WML) adding speech capability to these mark-up languages. While it is suggested to support the thin client architecture, applicability to mobile phones is limited due to the requirement of browsers' support for SALT. The need to manage multiple application versions, one for each device, is also observed as a drawback (Kondratova, 2004).

XHTML+Voice (X+V)

X+V is a proposed standard for multimodal markup. Designed for clients that support spoken and visual interaction, X+V technology furnishes traditional Web pages with further voice tags for input and output speech tasks. A goal is to support thin client, however in its current form this technology is most suited to the distributed client architecture. Use of VoiceXML has also been studied (Niklfeld et al., 2001).

Synchronized Multimedia Integration Language (SMIL)

SMIL is a relevant standard to managing media synchronization, an important multimodal problem. However, the standard is largely aimed at conventional browser technology that supports XML, DOM, and XHTML. Hence, it is restricted to browsers that may not be deployed to mobile phones.

User Interface Design

There is some work treating multimodal user interface design in mobility. A framework for multimodal interface design has been proposed in Baillie, Simon, Schatz, Wegscheider, & Anegg, 2005). The authors suggest the need to study user behavior in the natural environment. This follows the observation that user behavior differs considerably when comparing a controlled environment to the general freedom of use that mobile devices provide (Baillie & Schatz, 2005). A key feature of the method is a table designating the context in which the multimodal system is used and the users' preferred mode of interaction. Further work extends these concepts by proposing a tool to more fully analyze user behavior over an extended period of time (Salembier, Kahn, Calvet, Zouinar, & Relieu, 2005).

Given that there is little data on user behavior in a mobile environment, early design efforts will require careful consideration with users on how to maximize multimodal capabilities. Several key factors determine user behavior including response to error recovery, the course of action context, and properties of the implementation of each modality (Calvet, Julien Kahn, Pascal Salembier, & Zouinar, 2003). Further work shows that the speech interface appears to be the major contributor, when complementing a text or graphical interface, to the improvements bestowed through multimodality (Pavlovski et al., 2004b).

Summarizing the literature, the key features that a multimodal user interface design must address include the capability to dynamically alter modes, support for simultaneous composite output, and the ability to review the output request on one mode, using an alternative mode prior to actioning the request; for example, users who may dictate a message would prefer to see a transcript of the message prior to sending.

FUTURE TRENDS

There are several industry trends and future areas of research including the extension of multimodal solutions to accommodate haptic response, visual gesture recognition, lip reading, and speech technologies that support natural language understanding. Oviatt (2003) reviews several of these emerging technologies.

The idea of haptic response is already in use with present-day mobile phones, where mobile phones offer a silent mode that alerts the user with a vibration event. Such stimuli has been in use for some time and appear with early flight control systems that provide feedback to the pilots' control system indicating turbulent flight conditions. More recently, gaming controllers use such feedback. There is some recent work on the design of haptic response (Kaaresoja & Linjama, 2005) in mobile devices, while other work focuses on haptic input to control a mobile device (Oakley & O'Modhrain, 2005). The notion of extending applications in mobility with such haptic responses is relatively new. To extend such a capability to multimodal applications, access to native mobile services such as vibration is required to initiate a haptic response.

Lisetti et al. (2003) describe research progress on a multimodal emotion recognition system. They have developed a prototype that takes input from multiple modes (i.e., camera, microphone, wearable computer) and includes both subjective, as expressed by the user, and physiological components as inputs to its model. Lip reading technology, as an alternative interaction mode, may be used to improve speech recognition. Prior research suggests that bimodal audiovisual speech recognition can improve as much as 40% in accuracy (Chan, 2001), in comparison with using audio input alone. Additional work on the use of Natural Language understanding to augment the multimodal user experiences has been proposed and studied (Pavlovski et al., 2004b).

A final observation is required regarding security and privacy. Traditionally, voice communications are considered relatively secure, due to the transient nature of circuit establishment between parties and the real-time exchange of data between two peer entities. This is no longer the case for multimodal application, where the utterances are digitally recorded and stored. Furthermore, in order to improve the speech accuracy and fault resolution, there is a desire on the operators' part to store such data for a perceived user benefit. Of course the storage of personal conversations and requests for information raises several security and privacy issues that have yet to be addressed.

CONCLUSION

The literature has shown that multimodal user interfaces are able to provide a superior user experience when interacting with multimodal applications on mobile devices such as cellular phones. Furthermore, the applicability of multimodality is extending beyond traditional mobile Web applications to a greater range of domains. For instance, due to the plurality of user interface modes, multimodal systems provide a natural improvement for people with disabilities (Kvale et al., 2005). Although studies demonstrate that multimodal user interfaces provide a superior user experience on a mobile phone, it remains unclear whether such an improvement remains consistent for all types of applications.

Distributed client architectures are able to overcome several problems identified in the literature, however the practicality of deployment for mobile phones requires consideration. Browser-based multimodal architectures overcome deployment issues and provide an alternative and easier mechanism for allowing mobile devices to access services. In order to support multimodal applications that make use of three or more interaction modes, the use of thin client architecture is as yet unproven. In addition, further capabilities that may not be inherent within the devices are required, such as local composite management. Hence, distributed clients would seem a necessary choice to support these capabilities and appear most appropriate to address the needs of the increasing complexity due to several modes of interaction, particularly as the media synchronization of modes becomes more complex.

REFERENCES

Anegg, H., Dangl, T., & Jank, M. (2004). Multimodal interfaces in mobile devices—the MONA project. *Proceedings of the Workshop on Emerging Applications for Mobile and Wireless Access (www2004 Conference)*, New York.

Baillie, L., & Schatz, R. (2005). Exploring multimodality in the laboratory and the field. *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI 2005)* (pp. 100-107). Torento, Italy.

Baillie, L., Simon, R., Schatz, R., Wegscheider, F., & Anegg, H. (2005). Gathering requirements for multimodal mobile applications. *Proceedings of the 7th International Conference on Information Technology Interfaces (ITI 2005)*, Dubrovnik, Croatia (pp. 240-245).

Calvet, G., Julien Kahn, J., Pascal Salembier, P., & Zouinar, M. (2003). In the pocket: An empirical study of multimodal devices for mobile activities. *Proceedings of the HCI International Conference,* Crete, Greece, (pp. 309-313).

Chou, W., Shan, X., & Li, J. (2003). An architecture of wireless Web and dialogue system convergence for multimodal service interaction over converged networks. *Proceedings of COMPSAC 2003,* Dallas, TX, (p. 513).

Coutaz, J., Nigay, L., & Salber, D. (1993). Taxonomic issues for multimodal and multimedia interactive systems. *Proceedings of the Workshop on Multimodal Human-Computer Interaction (ERCIM '93)* (pp. 3-12). Nancy, France.

Hastie, H., Johnston, M., & Ehlen, P. (2002). Context-sensitive multimodal help. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02)*, Pittsburgh, PA, (p. 93).

Johnsen, M. H., & Kvale, K. (2005). Improving speech centric dialogue systems—The BRAGE project. *Proceedings of the Norsk Symposium on Signal Handling (NORSIG 2005)*, Stavanger, Norway.

Kaaresoja, T., & Linjama, J. (2005). Perception of short tactile pulses generated by a vibration motor in a mobile phone. *Proceedings of the 1st Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'05)* (pp. 471-472). Pisa, Italy.

Klante, P., Krösche, J., & Boll, S. (2004). AccesSights—A multimodal location-aware mobile tourist information system. *Proceedings of the 9th International Conference on Computers for Handicapped Persons (ICCHP 2004)* (pp. 287-294). Paris.

Kondratova, I. (2004). Speech-enabled mobile field applications. *Proceedings of the Conference on Internet and Multimedia Systems and Applications (IMSA 2004)*, Kauai, HI.

Kvale, K., & Warakagoda, N. (2005). A speech centric mobile multimodal service useful for dyslectics and aphasics. *Proceedings of EuroSpeech 2005,* Lisboa, Portugal, (pp. 461-464).

Kvale, K., Warakagoda, N., & Knudsen, J. (2003). Speech centric multimodal interfaces for mobile communication systems. *Telektronikk,* (Vol. 2, pp. 104-117). Temanummer: Spoken Language Technology.

Kvale, K., Warakagoda, N., & Kristiansen, M. (2005). Evaluation of a mobile multimodal service for disabled users. *Proceedings of the 2nd Nordic Conference on Multimodal Communication (Multimod Com '05)*, Gothenburg, Sweden.

Li J., Wong W., & Guo, W. (2004). Case study of a multimedia wireless system. *Proceedings of the IEEE International Conference on Multimedia and Expo, (ICME 2004)* (pp. 1815-1818). Taipei, Taiwan.

Niklfeld, G., Finan, R., & Pucher, M. (2001). Architecture for adaptive multimodal dialog system based on VoiceXML. *Proceedings of EuroSpeech 2001,* Aalborg, Denmark, (pp. 2341-2344).

Oakley, I., & O'Modhrain, S. (2005). Tilt to scroll: Evaluating a motion based vibrotactile mobile interface. *Proceedings of the 1st Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'05)* (pp. 40-49). Pisa, Italy.

Oviatt, S. L. (2003). Advances in robust multimodal interface design. *IEEE Computer Graphics and Applications, 23*(5), 62-68.

Pavlovski, C. J., Lai, J., & Mitchell, S. (2004a). Etiology of user experience with Natural Language speech. *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju Island, Korea, (pp. 951-954).

Pavlovski, C. J., Wood, D., Mitchell, S., & Jones, D. (2004b). Reference architecture for 3G thin client multimodal applications. *Proceedings of the International Symposium on Communications and Information Technologies (ISCIT 2004)*, Sapporo, Japan, (pp. 1192-1197).

Saeyor, S., Mukherjee, S., Uchiyama, K., & Ishizuka, M. (2003). A scripting language for multimodal presentation on mobile phones. *Proceedings of the 4th International Workshop on Intelligent Virtual Agents (IVA 2003)* (pp. 226-230). Kloster Irsee, Germany.

Salembier, P., Kahn, J., Calvet, G., Zouinar, M., & Relieu, M. (2005). Just follow me. Examining the use of a multimodal mobile device in natural settings. *Proceedings of the HCI International Conference,* Las Vegas, NV.

Sinha, A. K., & Landay, J. A. (2002). Embarking on multimodal interface design. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI 2002)* (pp. 355-360), Pittsburgh, PA.

W3C (World Wide Web Consortium). (2003a, January 8). *Multimodal interaction requirements.* Retrieved from http://www.w3.org/TR/mmi-reqs/

W3C (World Wide Web Consortium). (2003b, May 6). *Multimodal interaction framework.* Retrieved from http://www.w3.org/TR/mmi-framework/

KEY TERMS

Automated Speech Recognition (ASR): The use of computer processing to automatically translate spoken words into a text string.

Directed Dialog: Speech interface where the user receives system prompts that direct the user on available options that usually require a response with simple spoken words.

Fission: The decomposition of output from a multimodal application, to be distributed to the different modalities.

Fusion: The composition of input from multiple modalities into one input request to the multimodal application.

Multi-Call Supplementary Service: Enables several simultaneous circuit-switched calls to the same device, each call using its own dedicated bearer. Only one circuit-switched bearer can be used for speech at any one time.

Multimodal Interaction (MMI): W3C term for a class of human to computer interaction that involves multiple modes of user interaction, such as speech, writing, text, and haptics.

Text To Speech (TTS): The use of computer processing to transform text into spoken words.

Modular Sensory System for Robotics and Human-Machine Interaction Based on Optoelectronic Components

Milan Kvasnica
Tomas Bata University, Zlin, Czech Republic

INTRODUCTION

Presented here is a new unified modular sensory system. The subject of the article is the sampling and information processing used in the conversion of a 2-D CCD array image into three axial and three angular displacement values. The CCD array image consists of four light spots produced by four light beams (planes) from laser diodes. These light beams (planes) form the edges (faces) of a pyramidal shape, with the 2-D CCD array forming its base and the origin of the laser sources forming its apex. The algorithm for the computation of the location and orientation is based on the inverse transformation of the final trapezoidal light spots position, related to the original square light spots position on the 2-D CCD array. This algorithm determines the relative location and orientation of a floating 2-D coordinate system (corresponding to the 2-D CCD array) against a fixed 3-D coordinate system (corresponding to the apex of the pyramidal shape). The modular design presented here enables easy customizing of this sensory system for a wide variety of applications. Various combinations of the modular components enable tailoring of the sensory system properties for applications such as:

- portable modular system for the six-component dynamic measurement in general anisotropic construction in 3-D space,
- detection of microelastic or macroelastic deformation,
- six-DOF force-torque sensors of various properties,
- active compliant links,
- haptic interface,
- multi-DOF hand controllers,
- signature scanners for banking,
- keyboards for blind people,
- tactile sensors,
- range-incline finders-positioners,
- chaser systems,
- accelerometers,
- dynamic weighing, and
- artificial limbs.

In general, this modular design concept allows:

- maximization of service life because of ease of repair and the use of universal modular components for various types of sensors;
- environmentally friendly design, because the modular components are recyclable; and
- customization for a wide variety of design requirements.

Examples include various levels of resolution and operating frequency, enhanced demands for safety and reliability in space robotics, operation in dangerous areas and medical use with supporting self-checking and self-correcting algorithms, and low-cost design for manufacturing. In conclusion, regarding industrial relevance, many fields could benefit from the use of such a sensory system: robotics, telerobotics, rehabilitation robotics, intelligent automation, manufacturing and materials handling, automotive, marine and aerospace industry, medicine, ergonomics, safety accident prevention, defense, and banking.

The function of these sensors is based on the six-DOF system for the scanning of axial shifting and angular displacement. This simple construction enables low-cost customization, according to the demanded properties by means of the modular sensory system consisting of the following basic modules:

A: Stiff module of two flanges connected by means of microelastic deformable medium;
B: Compliant module of two flanges connected by means of macroelastic deformable medium;
C: The module of square CCD elements;
D: The module of the insertion flange with basic light sources configuration and focusing optics;
E: The module of the insertion flange with auxiliary light sources configuration and focusing optics;
F: The module of the plane focusing screen;
G: The module of forming focusing screen;
H: The module of the optical member for the magnifying or reduction of the light spots configuration;

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

I: The module of switchable muff coupling for changing the scanning mode for the micromovement and the macromovement-active compliance; and

J: The module for the preprocessing of scanned light spots configuration.

The problem of the customization of six-DOF sensory systems according to the enhanced accuracy and operating frequency of scanning of the six-DOF information is possible to improve by means of the modules:

K: The module of insertion flange with the configuration of light sources with strip diaphragms, creating the light planes with strip light spots;

M: The module of the single or segmented linear or annular CCD or PSD elements with higher operating frequency; and

N: The module of two, parallel working, concentric CCD annulars with higher reliability.

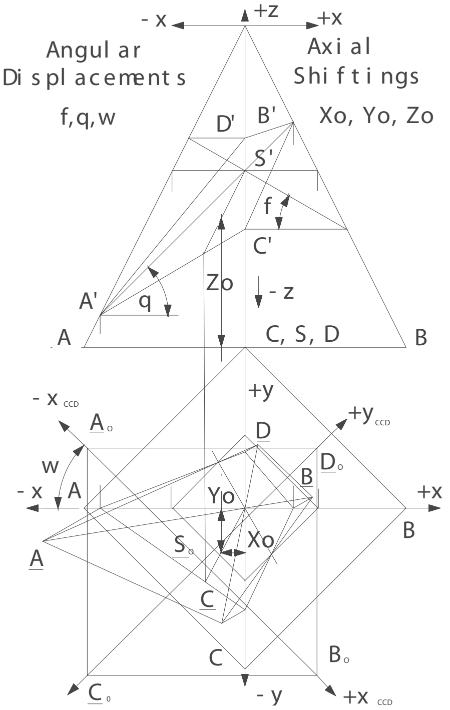
BACKGROUND

The results of our research program concerns the task of how to use a robot to imitate the human activity, as for example inserting a peg in a hole by means of the six-component force-torque sensor. The first robotic systems able to manage this task used a six-component strain gage wrist sensor developed at Stanford University, MIT, and C. S. Draper Laboratory (1960-1970). Using a robot to imitate a human activity in this way marked a historic step in human creative thinking. Recently, several implementations of sensory systems using optical imaging of six-DOF information on a CCD or PSD array have been proposed. An example is the project NASA in the cooperation with the DLR project ROTEX (1993).

SIX-COMPONENT FORCE-TORQUE SENSOR

The explanation of the activity of the majority of sensors described here is introduced with the six-component force-torque sensor (see Figures 1 and 2) composed from modules A,C,D,F,H, of the intelligent modular sensory system. Laser diodes 1 emit the light beams 2 creating the edges of a pyramid intersecting the plane of the square CCD element, here alternatively the focusing screen 8 with light spots 3. The unique light spots configuration changes under axial shifting and angular displacements between the inner flange 5 and the outer flange 6 connected by means of elastic deformable medium 7. An alternatively inserted optical member 9 (for the magnification of micromovement or the reduction of macromovement) projects the light spots configuration from the focusing screen onto the square CCD element 4. Four light beams simplify and enhance the accuracy of the algorithms for the evaluation of six-DOF information. The algorithms for the evaluation of three axial shiftings and three radial displacements are based on the inverse transformation of the final position of points A,B,C,D, related to the original basic position of points A_0,B_0,C_0,D_0,S_0 of the plane coordinate system x_{CCD}, y_{CCD} of the square CCD element (see Figures 1 and 2).

Figure 1. The approach of six-DOF scanning

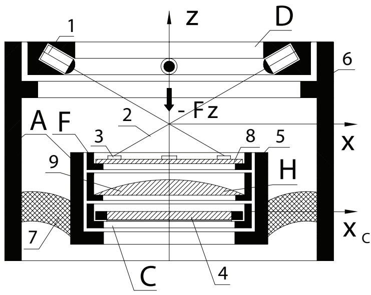


The information about axial shiftings caused by forces Fx, Fy, Fz and angular displacements caused by torques Mx, My, Mz are sampled and processed according to a calibration matrix. The intelligent modular sensory system enables us to compose in a customized way the various modifications of the multi-DOF force-torque sensors and compliant links for artificial arms or legs, range incline finders, hand controllers for wheelchairs, tactile sensors, keyboards for blind people, and handwriting scanners.

HUMAN ARTIFICIAL LIMBS

The effort to imitate by means of robot the human behavior of inserting a peg in a hole for the purposes of automatic assembly led to the development of the six-component force-torque sensor. For the scientist it is more satisfying to

Figure 2. Six-component force-torque sensor



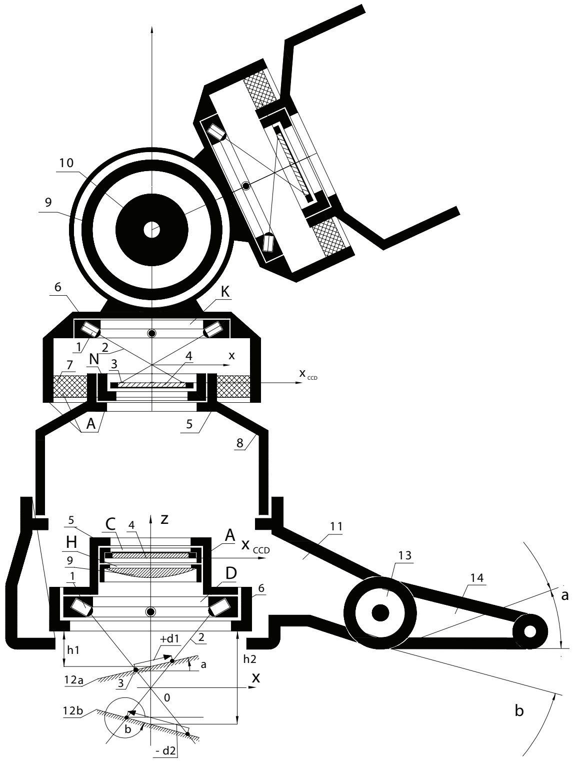
utilize such sensors to substitute for the missing limbs of the human body by an artificial limb of higher quality. Universal, low-cost, intelligent modular sensory systems enable us to evaluate a man's hand or leg dynamics while in motion. A part of the artificial leg consisting of the joint 10 connecting a shin with a foot 11 is depicted in Figure 3. The motion of the joint 11 is controlled by means of the six-DOF information gained from two six-component sensors. The joint's 10 drive transmission is switched by means of the coupling muff 9 in order to control the dynamics of the motion.

The six-component information about the leg's dynamics processed from two force-torque sensors enables us to use the drive power intelligently, even to convert the damping of the joint 10 motion for energy recuperation into the battery. The joint 13 connects the foot 11 with the toes part 14. The angular displacement (here for example a, b) of the joint 13 is used for accommodation to the ground's incline 12a, 12b, according to the information from the range-incline finder.

RANGE-INCLINE FINDER

The ground's incline under the artificial leg is scanned by means of the range-incline finder mounted in a heel (see Figure 3) consisting of the modules A, C, D, H. The light spots 3 from the light beams 2 on the ground 12a, 12b create the configuration scanned by the square CCD element. The processing of this information enables us to evaluate the incline of the ground in two perpendicular planes. Real-time algorithms are suitable for the single cheap microprocessor. An acoustic signal as indicator of the ground's incline helps the user to keep stability. The range-incline finder mounted on a wheelchair helps to keep the desired distance from a wall.

Figure 3. Six-component force-torque sensors mounted in artificial leg and the range-incline finder built in the heel



CUSTOMIZED DESIGN OF A DEXTEROUS HAND

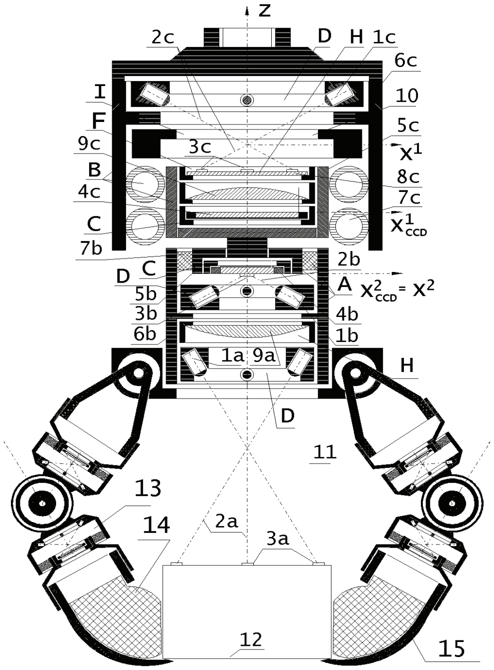
In rehabilitation robotics and in health care, many tasks occur frequently for example at the feeding of disabled people:

- the approaching of the artificial hand with the feeding utensil into the required position in front of a target object,
- the sequence of the operations until the time instant of the first contact with the target part of the body, and
- the inserting into a target part of a body.

Following this is the force-torque manipulation with a target object, with the aim here, for example, to load the food into the mouth and to protect the hurt.

Intelligent sensory systems for the solution of these tasks may be implemented instead of a missing part of a human hand or as a part of a robot's hand. In addition there is a possibility to evaluate the weight of gripped food in a dynamic way, while the robot's hand is in motion, in order to check the caloric limit.

Figure 4. Customized design of dexterous hand



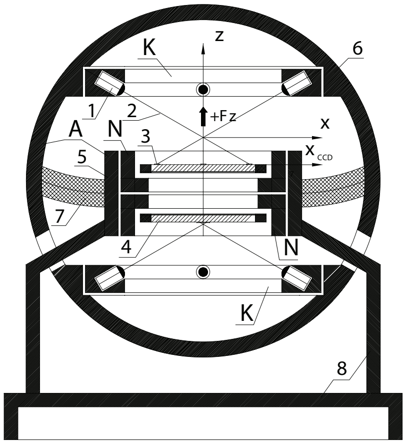
A simple solution of an universal dexterous hand consists of three sensory systems with two independently working CCDs (see Figure 4).

The first sensory system is the range-incline finder-positioner, composed of three modules C, D, H, alternatively working into the CCD element 4b. The range-incline finder-positioner consists of two pairs of mutual perpendicularly situated cross-light beams (planes) 2a radiated from the laser diodes 1a situated on the gripper. The configuration of the light spots (strips) 3a on the surface of the target object is projected by means of the zoom optical member 9a into the CCD element 4b. This multi-laser scanning equipment is used in the approach of the robot's gripper to the target and for simplifying some tasks in recognizing three-dimensional backgrounds.

The second sensory system is a six-component stiff force-torque sensor, composed of three modules A, C, D, alternatively working into the CCD element 4b. The laser diodes 1b fastened on the outer flange 6b radiate the light beams (planes) 2b against the CCD element 4b, fastened on the inner flange 5b. The unique light spots (strips) configuration 3b is changed under the force-torque acting between flanges 5b and 6b, both mutually connected by means of microelastic deformable medium 7b.

The third sensory system is the six-component active compliant link composed of six modules B, C, D, F, H, I, working into the CCD element 4c. The laser diodes 1c emits the light beams (planes) 2c against the focusing screen 8c. An optical member 9c mediates the reduction of the macro-movement of the light spots (strips) 3c. The unique light spots (strips) configuration 3c is changed under the force-torque acting between flanges 5c and 6c, connected by means of the active compliant medium 7c. An active compliance is solved by means of pneumatic, programmable switched, segmented hollow rubber annulars 7c. Alternative use of the six-component stiff force-torque sensor or the active compliant link is switched by means of coupling muff 10.

Figure 5. Multi-DOF hand controller

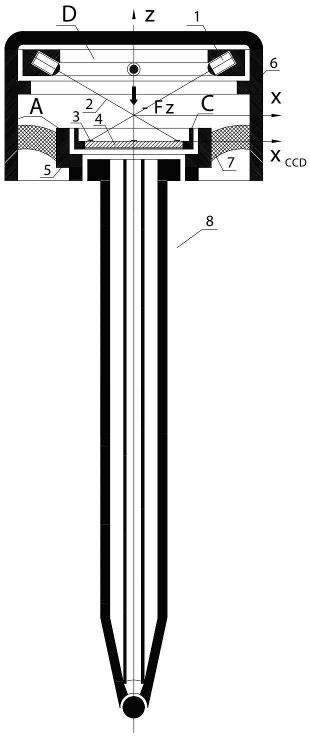


Dexterous fingers 13 are equipped by two force-torque sensors for the motion control of the finger's joints by means of the six-component information. The top of every finger is equipped by elastic rubber cushions 14 and by nails 15. The unified modular intelligent sensory system enables customized design for a wide variety of tasks in rehabilitation robotics.

HAND CONTROLLER

Efficiency in using a wheelchair depends on the user's effectiveness in communicating with the driving gear. A low-cost six degrees-of-freedom hand controller means for many users not luxury, but the possibility for personal autonomy in their daily activities. A multi-DOF hand controller is possible to use for the control of the feeding utensil combined with a simple mechanism. The multi degrees-of-freedom hand controller (low cost) or of enhanced reliability is depicted in Figure 5, under the influence of the acting force +Fz. This device consists of the module C of the square CCD element, or of the module N, for example in medical use of enhanced reliability for surgeons with two independently parallel working CCD annulars 4, fastened in mutually opposite directions in front of the (module D) modules K of the (light beams) light planes 2. The configuration of the (light beams) light planes 2 of the pyramid shape is radiated from the laser di-

Figure 6. Keyboard for blind people



odes 1 fastened on the outer flange 6. The configuration of light (beams) planes 2 creates in the plane of (square CCD elements) the CCD annulars of the configuration of light (spots) strips 3. The inner flange 5 is fastened on the stand 8 and connected by means of the elastic deformable coupling balks 7 with the outer flange 6. The design of the outer flange 6 is shaped for a human-hand-friendly form.

KEYBOARD FOR BLIND PEOPLE

Six-component force-torque sensors that make it possible to pass judgment about the heterogeneity of a man's hand dynamics, for example the handwriting of two different persons, may be used like a keyboard for blind people. Because of the lack of place for the six-component force-torque sensor between a nib and a penholder, the configuration seen in Figure 6 was used, where the inner flange 5 is put on the end of a penholder 8. The outer flange creates a steady mass. This handwriting scanner is possible to use as a keyboard for blind people in order to improve their communication with a computer. Another configuration of the handwriting scanner, where the six-component force-torque sensor is inserted between the writing plate 6 and the support 8 of the writing hand, is depicted in Figure 7. This device may be used as a signature scanner in banking.

Figure 7. Six-DOF signature scanner

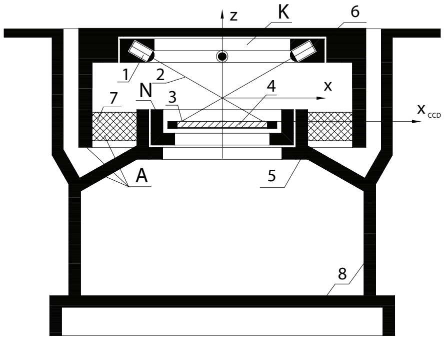
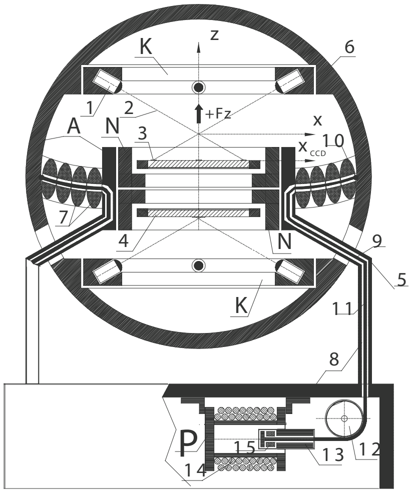


Figure 8. Six-DOF hand controller with haptic interface



HAND CONTROLLER WITH HAPTIC INTERFACE

Adding haptics to an interface can enhance human-computer interaction. A haptic interface enables software to utilize a new class of operations, including cursor-controlled feedback through dynamic force-torque responses from virtual objects displayed on screen. It also allows novel hardware interfaces, for example in the handling of fragile materials or assembly operations in telerobotics or in space robotics. In rehabilitation robotics, it is possible to use a multi-DOF hand controller with haptic feedback as a human-machine interface, for example for the control of a feeding utensil. A six-DOF hand controller (see Figure 8), with the same basic structure as shown in Figure 2, shows the inner flange 5 fastened on the stand 8 and connected by means of the elastic medium 7, with the outer flange 6, which is shaped

Figure 9. Six-DOF head controller with haptic interface

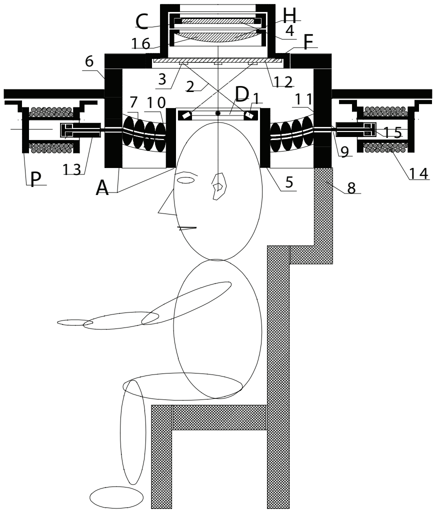
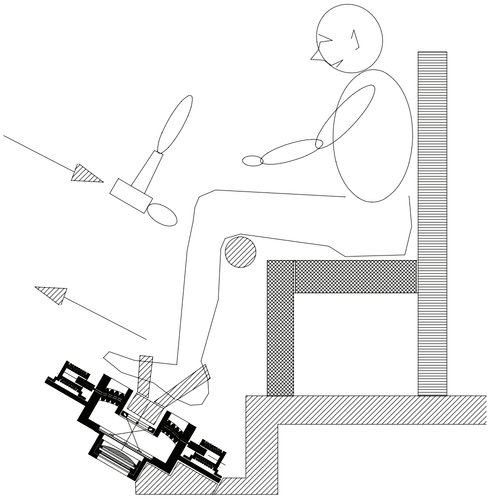


Figure 10. The examination of the patellar reflex response by means of the leg's dynamics movement scanner



to be grasped comfortably. One end of the Bowden cable 9 is fastened on the outer flange in the point 10 and inserted in the hole 11 in the inner flange 5 and in the stand 8. The other end of the cable passes through the elastic medium 7 and the wheel 12, and is fastened to the movable core 13 of electromagnetic actuator P. The pressure sensor 15 for the calibration of haptic forces is inserted between the foot of the cable 9 and the movable core 13.

HEAD CONTROLLER WITH FORCE INTERACTION

For 3-D tasks in space robotics, telerobotics, or surgery, a head controller can be useful. A six-DOF head controller with force feedback interface is depicted in Figure 9. The modular design allows this device to be constructed identically to the above hand controller, but uses larger parts, with the exception of the attachments of the inner and outer flanges. The inner flange 5 is fastened on the human head, and the outer flange is fastened to the stand 8 from the chair. The inner flange 5 is connected by means of the elastic medium 7 with the outer flange 6.

Another possibility of how to use the head controller with force interaction is depicted in Figure 10. There is an application for the examination of the patellar reflex response in neurology by means of the hit of the reflex hammer, connected by means of the force sensor with computer. The examination of the patellar reflex response using the leg's dynamics movement scanner is evaluated from the course of the dynamic characteristic by means of the computer. Another application is possible like the leg's controller.

SIX-DOF SENSORY SYSTEM FOR INTERACTIVE POSITIONING IN SURGERY

Some applications of multi-pod's parallel structures frequently need the six degrees-of-freedom (DOF) force torque feedback between the multi-pod's platform with a patient and a robot's effector at robotic-aided therapy. A six-DOF force-torque transducer is inserted between the multi-pod's platform with a patient and a stand (see Figure 11). The difference between the actual, measured, and computed position of both platforms changes during the manipulation. The first difference arises under the influence of interpolation of the positioning algorithm, elasticity of materials, temperature dilatation, inaccuracy of incremental sensor, and servo drivers. The second difference is caused by the functional elasticity of the body of an inserted force-torque transducer, and both have an additive course.

The multi-pod's motion sensing system is derived from the pyramid modular sensory system for the sampling of information about six-DOF displacements (see Figure 12). The subject of this modular sensory system is based on the conversion of four 2-D PSD (CCD) array images into three axial shifting and three angular displacement values. Every 2-D PSD (CCD) array image consists of one light spot produced by light rays from four laser sources. These light rays form the edges of a pyramidal shape, with four 2-D PSD (CCD) arrays forming its base in the vertex of the square shape. An origin of the laser sources 9 is forming its apex P.

The sampling of the configuration between two platforms by means of two universal modules is depicted in Figure

Figure 11. Multi-pod's parallel structure combined with the six-DOF force-torque transducer

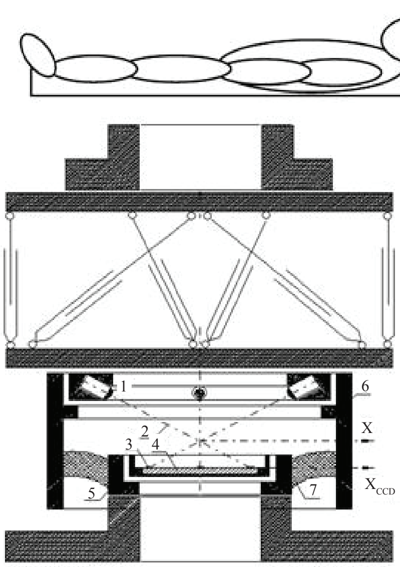
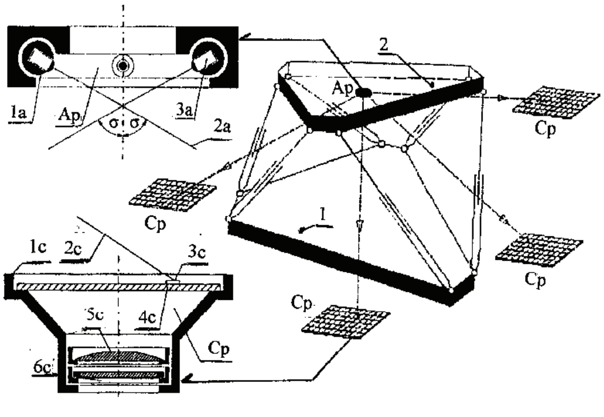


Figure 12. Multi-pod's parallel structure equipped by the module Ap of four lasers and four modules Cp for direct sampling



11. Here the position of movable platform 2 is sampled by the module Ap of four laser 3a radiated light rays 2a with presetting control 1a of the angle 2σ against four modules Cp mounted on the stand 1. Modules Cp are situated in the corners of a square shape on the multi-pod's stand. The module Cp for direct sampling of measured values fluctuation of the light spot 3c position consists of the 2-D PSD (CCD) array 6c with focusing optics 5c and the flange 1c. The light spot 3c from the laser light ray 2c is imagined on the translucent screen 4c.

Figure 13. The module M of segmented linear CCD or PSD arrays intersected by structured light planes for enhanced scanning frequency

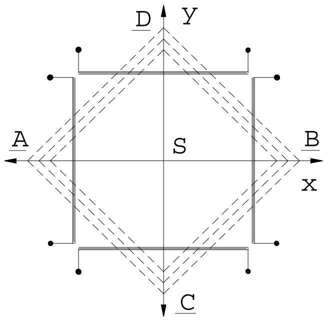
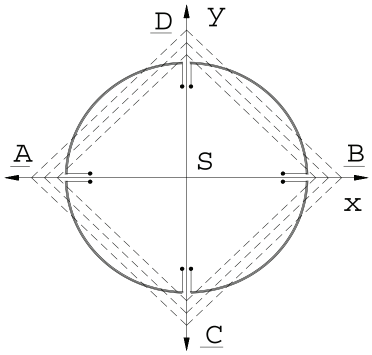


Figure 14. The module M of segmented annular CCD or PSD arrays intersected by structured light planes for enhanced scanning frequency

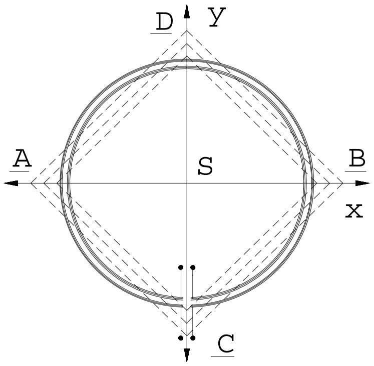


THE ENHANCEMENT OF THE SAMPLING FREQUENCY

The problem of the customization of six-DOF sensory systems according to the enhanced accuracy and operating frequency of scanning of the six-DOF information is possible to improve by means of the modules:

K: Module of insertion flange with the configuration of light sources with strip diaphragms, creating the light planes with strip light spots;
M: Module of the single or segmented linear or annular CCD or PSD elements with higher operating frequency; and
N: Module of two, parallel working, concentric CCD annulars with enhanced reliability.

Figure 15. The module N of two parallel working concentric annular CCD or PSD arrays intersected by structured light planes for enhanced scanning frequency and enhanced reliability



CONCLUSION

The modular design presented here enables easy customizing for a wide variety of applications. Various combinations of the modular components enable tailoring of the sensory system properties, including the use of the haptic interface for applications such as detection of microelastic or macroelastic deformation, active compliant links, multi-DOF hand controllers, signature scanners, keyboards for blind people, tactile sensors, and range finders-positioners. In general, this modular design concept allows maximization of service life because of ease of repair and the use of modular components for various types of sensors, and customization for a wide variety of design requirements, for example, various levels of resolution and operating frequency, enhanced demands for safety and reliability in space robotics and medical use, and low-cost design for manufacturing.

In conclusion, many fields could benefit from the use of such a sensory system: robotics, telerobotics, measurements in engineering constructions, intelligent automation, automotive and aerospace industry, medicine, defense, and banking.

REFERENCES

Hirzinger, G., Dietrich, J., Gombert, J., Heindl, J., Landzettel, K., & Schott, J. (1992). The sensory and telerobotic aspects of space robot technology experiment ROTEX. *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space,* Toulouse, France.

Kvasnica, M. (1993, September). Fast sensory system for the scanning of six-component axial shiftings and radial displacements. *Proceedings of the 3rd IMEKO International Symposium on Measurement and Control in Robotics,* Torino, Italy.

Kvasnica, M. (1997, July). Flexible sensory brick-box concept for automated production and man-machine interface. *Preprints and Proceedings of the ICIMS-NOE International Conference on Life Cycle Approaches to Production Systems, Management, Control, Supervision,* Budapest, Hungary.

Kvasnica, M. (1999, July). Modular force-torque transducers for rehabilitation robotics. *Proceedings of the IEEE International Conference on Rehabilitation Robotics,* Stanford, CA.

Kvasnica, M. (2001, April). A six-DOF modular sensory system with haptic interface for rehabilitation robotics. *Proceedings of ICORR'2001, the 7th International Conference on Rehabilitation Robotics,* Paris-Evry, France.

Kvasnica, M. (2001). Algorithm for computing of information about six-DOF motion in 3-D space sampled by 2-D CCD array. *Proceedings of the 7th World Multi-Conference* (SCI2001-ISAS, vol. XV, Industrial Systems, Part II), Orlando, FL.

Kvasnica, M. (2002). Six DOF measurements in robotics, engineering constructions and space control. *Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics* (*SCI 2002-ISAS 2002*, ext. vol. XX), Orlando, FL.

Kvasnica, M. (2003). Six-DOF sensory system for interactive positioning and motion control in rehabilitation robotics. *International Journal of Human-Friendly Welfare Robotic Systems, 4*(3).

Kvasnica, M. (2004, September). Six-DOF force-torque transducer for the wheelchair control by means of the body motion. *Proceedings of the 2nd International Conference on Smart Homes and Health Telematics (ICOST 2004)*. Singapore.

Kvasnica, M. (2005). Assistive technologies for man-machine interface and applications in education and robotics. *International Journal of Human-Friendly Welfare Robotic Systems, 6*(3).

Kvasnica, M. (2005, October 19-22). The accuracy of six degrees of freedom sensory systems. *Proceedings of the 16th International DAAAM Symposium on Intelligent Manufacturing and Automation,* University of Rijeka Opatia, Croatia.

Kvasnica, M. (2005, September 26-29). The analysis of direct transformation for floating image coordinates in 3-D coordinate frame of the six-DOF sensory system. *Proceedings of the 6th International Conference on Mechatronics, Robotics and Biomechanics 2005,* Třešť, Czech Republic.

Kvasnica, M., & Vašek, V. (2004, May). Mechatronics on the human-robot interface for assistive technologies and for the six-DOF measurements systems. *Proceedings of the 7th*

International Symposium on Topical Questions of Teaching Mechatronics, Rackova Dolina, Slovakia.

Kvasnica, M., & Vašek, V. (2005). Force-torque wheelchair control by the body motion. *Sborník a přednáška na Konferenci SSKI s mezinárodní účastí Kybernetika a informatika Dolný Kubín.*

Kvasnica, M., & Van der Loos, M. (2000). Six-DOF modular sensory system with haptic interaction for robotics and human-machine interaction. *Proceedings of the World Automation Congress (WAC 2000)*, Maui, HI.

KEY TERMS

Deutsche Luftraum (DLR): German aerospace establishment.

Hand Controller: Joystick.

Human-Machine Interface: Mutual interaction between a man and a machine.

Laser Diode: Light-emitting source based on semiconductors.

Position-Sensitive Device (PSD):

Robot Technology Experiment (ROTEX): NASA project of the Aerospace Laboratory launched in 1993.

Monitoring and Tracking Moving Objects in Mobile Environments

Dragan Stojanović
University of Nis, Serbia

Slobodanka Djordjevic-Kajan
University of Nis, Serbia

Apostolos N. Papadopoulos
Aristotle University, Greece

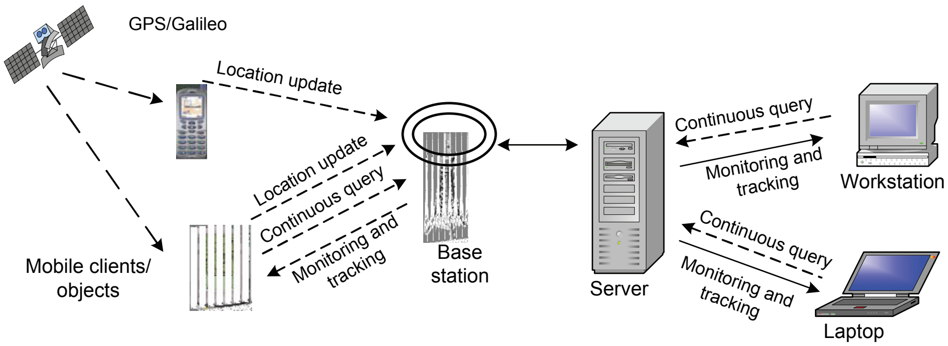
Alexandros Nanopoulos
Aristotle University, Greece

INTRODUCTION

Advances in mobile and ubiquitous computing, wireless communications and mobile positioning have given a rise to a new class of mobile information systems and services, called location-based services (LBS) (Schiller & Voisard, 2004). Such services, like fleet management, cargo tracking, childcare, tourist services, transport management, traffic control and digital battlefield rely on the tracking of the continuously changing positions of entire populations of moving objects. LBSs are becoming ubiquitous. A traffic service may inform its users about traffic jams, traffic accidents and weather situations that are expected to be of relevance to the service user. A friend finder service may inform each user about the current whereabouts of friends. Other services may monitor and track the positions of emergency vehicles, police cars, security personnel, hazardous materials, or public transport. A more advanced location-based game service may allow a group of users to play and try to surround and catch "hostile" players. Monitoring and tracking moving objects require continuously registering the positions of mobile objects, and at any instance in time to know whether those objects are within the specified area, or a specified distance from known mobile/static objects, or which are its *k* nearest neighbors (*k*-NN). To provide monitoring and tracking of moving objects in mobile application environments, it is highly desirable and sometimes critical for the service efficiency to provide accurate results to these requests and update them in real time, whenever moving objects enter or exit the regions of interest, or become the closest neighbors to the objects of interest.

Monitoring and tracking LBS applications require database and application support to model and manage moving objects in both database and application domains. Such services must also provide efficient processing of continuous queries over moving objects. In contrast to regular queries that are evaluated only once, a continuous query remains active over a period of time. At any time there will be a number of continuous queries simultaneously running at the server. Each of these queries needs to be periodically re-evaluated as the objects and/or queries move. A major challenge for this problem is how to provide efficient processing of continuous queries with respect of CPU time, I/O time and network bandwidth utilization. The architecture of the monitoring and tracking LBS system is given in Figure 1.

Figure 1. The architecture of monitoring and tracking LBS



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

BACKGROUND

Monitoring and tracking moving objects in mobile environments by processing continuous queries over moving objects is an active area of research, resulting in the proposal of many query processing methods, techniques and indexing schemes. One of the challenges in monitoring and tracking LBS development is how to handle different types of queries in a mobile environment, where both queries and objects can be moving. Different types of location dependent queries are significant for the monitoring and tracking purposes, such as range queries, *k*-nearest neighbor (*k*-NN) queries, reverse neighbor queries, distance joins, closest pair queries and skyline queries. The most important type of query for the purpose of monitoring and tracking moving objects is the range query. The range may represents a user selected area, a map window, a polygonal feature, a part of the road segment or an area specified by the distance from a reference point. The map window of the LBS client application represents the simplest continuous range query that must be supported in monitoring and tracking LBS application. Using such a query, up-to-date information about moving/stationary objects in a user's surrounding is continuously represented in the map window, as he/she, as well as objects of interest, move.

Continuous query processing in a location-aware environment is an active area of research, resulting in the proposal of many query processing methods, techniques and indexing schemes. In Prabhakar et al. (2002), velocity constrained indexing and query indexing (Q-index) has been proposed for efficient evaluation of stationary continuous range queries. According to the proposed method, in-memory data structures and algorithms are developed and presented (Kalashnikov et al., 2004). By indexing queries, and not moving objects, the Q-index method avoids frequent updates of the index structure and thus expensive maintenance of this structure. The MQM method presented (Cai et al., 2004) focuses on stationary continuous range queries. It is based on partitioning the query space into rectangular sub-domains, and the assignment of the resident domain to each moving object. A moving object is aware only of the range queries intersecting its resident domain, and reports its current location to the server only if it crosses the boundary of any of these queries. Gedik and Liu (2004) propose a method and a system for distributed query processing, called Mobieyes. Mobieyes ships some part of the query processing to the mobile clients while the server mainly acts as a mediator between moving objects. The method tries to reduce the load on the server and save communication costs between moving objects and the server. In the paper by Gedik et al. (2004), the authors propose a scheme called motion adaptive indexing (MAI), which enables optimization of continuous query evaluation according to the dynamic motion behavior of the objects. They use the concept of motion sensitive bounding boxes (MSB) to model and index both moving objects and moving queries. Mokbel et al. (2004a) present SINA, a server-side method based on shared execution and incremental evaluation of continuous range and *k*-NN queries. Shared execution is achieved by implementing query evaluation as a spatial join between the moving objects and the queries. Incremental evaluation means that the query processing system produces only the positive or negative updates of the previously reported answer, not the complete answer for every evaluation of the query. Both the object and query indexes are implemented as disk-based regular grids. The same authors in Mokbel et al. (2004b) present a continuous query processor that extends a relational database management system and a data stream management system, to support efficient continuous query processing of spatiotemporal streams. They implemented the proposed query processor inside the PLACE (pervasive location-aware computing environments), scaleable location-aware database server. Based on the authors' previous work, the proposed continuous query processor provides incremental evaluation of continuous queries, shared and scaleable execution of a set of concurrent continuous queries and integration of data streaming management and semantics to support location-aware environments.

Hu et al. (2005) propose a generic framework for monitoring continuous spatial queries over moving objects, both range and *k*-NN queries. The work of Tao et al. (2002) focuses on continuous *k*-NN query evaluation, for moving queries over stationary objects. They propose an algorithm for precalculating *k* nearest neighbors with a line segment representing the continuous motion of the object.

Koudas et al. (2004), propose a method, called DISC, for approximate processing of *k*-NN queries over streams of multidimensional points, where the returned kth NN point is further than the actual *k*th NN point, within a specified distance threshold. Yu et al. (2005) describe a method for continuous monitoring of *k*-NN queries. Their method uses a main memory grid as an index structure and utilizes two algorithms using grid indices. The first one is based on object indexing, and the second is based on query indexing. Each *k*-NN query is evaluated for the first time by a two-step NN search technique, and is further re-evaluated every *T* time units. The initial step visits the cells in squares around the cell covering the query point until *k* objects are found. The second step refines the search by examination the cells outside the examined squares in order to determine the actual *k*-NN set of objects and remove false candidates appearing in the initial step.

Xiong et al. (2005) propose the method SEA-CNN for monitoring changes in the *k*-NN set, assuming that the initial result of a query is available. Objects are indexed by a regular grid on the disk. The method defines the answer region of a *k*-NN query as the circle centered at the query point with radius the distance to the current *k*-th NN, and the cells that intersect the answer region hold information about it. When the data or query objects update their locations, the SEA-CNN algorithm is used to determine the new *k*-NN set based on the new circular search region.

MONITORING AND TRACKING MOVING OBJECTS

The methodology for processing continuous range queries in a mobile environment is developed as a part of ARGONAUT, a service platform for moving object data management (Predic et al., 2005). We base our approach on the application scenario appropriate in LBS for monitoring and tracking moving objects. In this scenario, users have wireless devices (e.g., mobile phones or PDAs) that are online via some form of wireless communication network. We assume that users can obtain their positions using global positioning system (GPS) technology. A setting is assumed in which a central database at the LBS server stores a representation of each moving object's current position. Each moving object stores locally the representation of its current position assumed by the server. Then, an object updates the database whenever the deviation between its actual position (as obtained from a GPS device) and the local copy of the position that the central database assumes exceeds the uncertainty threshold (Civilis et al., 2005).

In most real-life monitoring and tracking applications, the object movement is constrained by an underlying spatial network, that is, objects can not move freely in space, and their position must satisfy the network constraints. Network connectivity is usually modelled by a graph representation, comprising a set of nodes (intersections) and a set of edges (segments). In order to perform tracking with as few updates as possible, the LBS server must match the position of the moving object sent in the form of (*X*, *Y*) coordinates to the underlying network segment. Map matching is a technique that positions an object on a network segment, at some distance from the start of that segment, based on location information from a GPS device. Knowing the moving object's speed (also sent by the moving object along with position) and the time of position update, the server can predict the current and the near future position of the object until the next junction (node) assuming that it moves at a constant speed (Civilis et al., 2005). Reduction of updates reduces communication between clients and the server and server side update and query processing.

The ARGONAUT methodology employs an incremental continuous query evaluation paradigm similar to Mokbel et al. (2004a, 2004b). Thus the server reports to the clients only the changes of the answer from the last evaluation time of their continuous queries. This significantly saves the network bandwidth by limiting the amount of transmitted data to the updates of the answer only rather than the whole query answer. Two types of updates are distinguished: positive updates and negative updates. The positive/negative update indicates that a certain object needs to be added/removed to/from the query answer.

As mentioned in the previous section, most query processing methods index either objects, queries or both in main memory. Thereby, efficient query processing and minimization of CPU time are achieved, both necessary for real-time applications. Since the object movement is constrained by the underlying transportation network, the methodology maintains two representations of the network data. The first representation organizes network segments according to Euclidean distance in a main memory R*-tree index structure (Beckmann et al., 1990), that stores MBR (minimal bounding rectangle) representation of segments belonging to a spatial network.

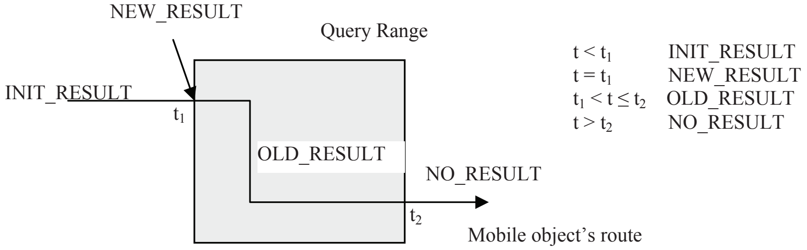
The graph representation of a spatial network is maintained by the network connectivity table (NCT) data structure, which stores information about the connectivity of network segments. Both network representations are interconnected, that is, there is a reference from R*-tree segment representation to the corresponding NCT segment representation and vise versa. For each segment, two lists are maintained: the list of objects IDs (*Olist*) and queries IDs (*Qlist*) that move along or reside on that network segment. Both segment representations have pointers to these lists.

A moving object ID is inserted in the *Olist* that corresponds to the network segment at which such object resides or currently moves according to map matching technique. The moving/stationary query is determined by the reference object and query condition (range, distance, *k*-NN, etc.). The stationary/moving query is inserted in each *Qlist* that corresponds to the network segments whose MBR representation is intersected by the query range (for Euclidean metric space), or which are in specified network distance from the reference object of a query (for network metric space).

By inserting the objects IDs and queries IDs in appropriate lists related to network segments, the matching between moving objects and moving/stationary queries according to their spatial relations is performed. This matching provides the filter step of continuous query processing. The ARGONAUT query processing methodology involves an additional step in the query processing strategy. The pre-refinement step is performed to further refine the initial query result set regarding object and query temporal information, as well as exact locations and query ranges. The pre-refinement step creates data structures in main memory to support incremental refinement steps.

The two tables and associated lists are created in main memory by the pre-refinement step. *Continuous query table* (CQT) stores information related to continuous queries. A CQT entry is described as (*QID*, *OID*, *range*, *resultSet*) and stores information related to a continuous query. The table is indexed on the *QID* attribute, which is the unique query identifier. *OID* is the identifier of the reference object

Figure 2. Changing the status of a moving object in a continuous query result



of the query; *range* defines the shape of the spatial query range around the reference query object. *resultSet* is the query result set obtained by the pre-refinement step with additional, temporal information about satisfaction of a query. The result set is a list of elements defined as (*RID*, *OID*, *resPeriod*, *status*), where *RID* is the result identifier, *OID* is the unique identifier of the object, which is the result of the query during period *resPeriod*, while its status in the query result is described by the *status* attribute. The values of the *status* attribute are INIT_RESULT, NEW_RESULT, OLD_RESULT, NO_RESULT. In the simplified case, when the resulting period is single time period, the resulting object, during its motion and/or query motion, change all status values sequentially (Figure 2).

For each moving object in the system, an in-memory mobile object table (MOT) is created and maintained. The moving object entry is described as (*OID*, *loc*, *time*, *speed*, *IDseg*, *querySet*), where *OID* is the unique moving object identifier, *loc* is the last received location update, *time* is the timestamp of the location update and *speed* is the last received speed. The *IDseg* attribute *is* the identifier of the current network segment. *querySet* attribute represents the list of queries in which such object participates, either in a query result, or as a reference object of a query.

The pre-refinement algorithm examines the set of moving objects obtained by the filter step, and creates *CQT* and *MOT* along with corresponding lists. For each query in the set, the pre-refinement algorithm generates the result of the query along with the validity period of the resulting object in that result. For each moving object in the system, the pre-refinement algorithm generates the set of queries in whose results it participates.

The incremental evaluation (refinement step) is performed periodically and evaluates the validity periods of query results in regard to current time. It determines the incremental result containing only the changes of the results from the last evaluation, that is, only positive and negative updates. This enables fast query evaluation and saves the network bandwidth by limiting the amount of data transmitted to the client.

The most challenging part of the query processing methodology in such real-rime settings is to update the data and index structures upon receiving location/speed/segment update of the moving object. When a moving object sends its location/speed update to the server, when its actual location differs from the server-predicted location by the specified threshold, the methodology needs to update the validity periods for all queries in whose result the particular object participates. If the moving object is the reference object of the query (-ies), the validity periods for all resulting objects must be updated accordingly. For the case of range queries, this update is represented by simple algebraic expressions. But, when a moving object changes its underlying network segment, the *OID* of the object must be deleted from previous *Olist* and entered in the new one associated to the new network segment. Thereby, the object can participate in the query results of the new set of continuous queries (new *Qlist*), and must be added in their *resulSets*, along with updating of its *querySet* accordingly. The complete algorithms for the query processing methodology and performance study for moving objects that know its destination in advance and thus their route/path in the network (public transport, fleet management, etc.) are given in Stojanovic et al. (2005). The improvement of such algorithms and experimental evaluation of our continuous query processor for objects moving freely on the network is presented in Stojanovic et al. (2006). The experimental results show that the performance of the ARGONAUT query processing methodology is satisfactory for real-world settings in LBS applications for monitoring and tracking moving objects. We plan to further extend and improve our methodology for query processing of continuous *k*-NN queries.

FUTURE TRENDS

The application of continuous query processing techniques to the monitoring and tracking of moving objects is a very active area for both research and practice. One of the future directions in the field represents continuous query processing over moving objects regarding their trajectories as data streams and incorporating data streams techniques in such process. Also, with the advancement of mobile computing/communication devices, distributed and mobile query processing techniques become more important. Such techniques ship some part of query processing to the moving objects, which have the computational and storage capabilities to perform some part of the query processing algorithms. The next step is to distribute data management, as well as query processing functionality required for monitoring and tracking moving objects to the mobile peer-to-peer networks, where moving objects communicate with each other via short-range wireless transmission. In such environments, moving objects serve as both clients that monitor and track other objects, and routers of queries and answers on behalf of other clients.

CONCLUSION

The wide-spread of smart cellular phones, personal digital assistants and GPS technology enables a mobile environment where objects are continuously moving and are aware of their locations. During the past few years, we have witnessed the emergence of location-based services field as a novel research area, investigating interesting problems and developing real-life applications and services related to monitoring and tracking moving objects.

Services such as these rely on the efficient management of moving objects' location and processing of continuous queries. Since most of these services have real-time constraints and are mission critical, it is of great importance to effectively process continuous queries and generate incremental results, which must be delivered to clients across a wireless/wired communication channel.

The field provides many attractive topics for both theoretical and engineering achievements and it is expected to be one of the key fields in mobile and ubiquitous computing research for the years ahead.

REFERENCES

Beckmann, N., Kriegel, H.-P., Schneider, R., & Seeger., B. (1990). The R*-tree: An efficient and robust access method for points and rectangles. In *SIGMOD* (pp. 322-331).

Brinkhoff, T. (2002). A framework for generating network-based moving objects. *GeoInformatica, 6*(2), 153-180.

Cai, Y., Hua, K., & Cao, G. (2004). Processing range-monitoring queries on heterogeneous mobile objects. In *Mobile Data Management* (pp. 27-38).

Civilis, A., Jensen, C., & Pakalnis, S. (2005). Techniques for efficient road-network-based tracking of moving objects. *IEEE Transactions on Knowledge and Data Engineering*, *17*(5), 698-712.

de Almeida, V., & Guting, R. (2005). Indexing the trajectories of moving objects in networks. *GeoInformatica, 9*(1), 33-60.

Frentzos, E. (2003). Indexing objects moving on fixed networks. In *8th International Symposium on Spatial and Temporal Databases* (pp. 289-305).

Gedik, B., & Liu, L. (2004). Mobieyes: Distributed processing of continuously moving queries on moving objects in a mobile system. In *Extensible Database Technology (EDBT)* (pp. 67-87).

Gedik, B., Wu, K., Yu, P., & Liu, L. (2004). Motion adaptive indexing for moving continual queries over moving objects. In *CIKM* (pp. 427-436).

Hu, H., Xu, J., & Lee, D. (2005). A generic framework for monitoring continuous spatial queries over moving objects. In *ACM SIGMOD Conference*.

Kalashnikov, D., Prabhakar, S., & Hambrusch, S. (2004). Main memory evaluation of monitoring queries over moving objects. *Distributed and Parallel Databases*, *15*(2), 117-135.

Koudas, N., Ooi, B. C., Tan, K. L., & Zhang, R. (2004). Approximate NN queries on streams with guaranteed error/performance bounds. In *Proceedings of the 30th International Conference on Very Large Databases (VLDB'04)* (pp. 804-815).

Mokbel, M., Xiong, X., & Aref, W. (2004). SINA: Scalable incremental processing of continuous queries in spatio-temporal databases. In *ACM SIGMOD Conference* (pp. 623-634).

Mokbel, M., Xiong, X., Hammad, M., & Aref, W. (2004). Continuous query processing of spatiotemporal data streams in PLACE. In *Proceedings of the Second Workshop on Spatio-Temporal Database Management* (pp. 57-64).

Prabhakar, S., Xia, Y., Kalashnikov, D., Aref, W., & Hambrusch, S. (2002). Query indexing and velocity constrained indexing scalable techniques for continuous queries on moving objects. *IEEE Transaction on Computers, Special Issue on DBMS and Mobile Computing*, *51*(10), 1124-1140.

Predic, B., & Stojanovic, D. (2005). A framework for handling mobile objects in location based services. In *AGILE* (pp. 419-427).

Schiller, J., & Voisard, A. (Eds.). (2004). *Location-based services*. San Francisco: Morgan Kaufmann/Elsevier.

Stojanovic, D., Djordjevic-Kajan, S., & Predic, B. (2005, December 15-16). Incremental evaluation of continuous range queries over objects moving on known network paths. In *Fifth International Workshop on Web and Wireless Geographical Information Systems* (LNCS 3833, pp. 168-182). Lausanne, Switzerland: Springer-Verlag.

Stojanovic, D., Djordjevic-Kajan, S., Papadopoulos A.N., & Nanopoulos, A. (2006). Continuous range query processing for network constrained mobile objects. *Proceedings of the 8th International Conference on Enterprise Information Systems (ICEIS 2006)* (pp. 63-70).

Tao, Y., Papadias, D., & Shen, Q. (2002). Continuous nearest neighbor search. In *Proceedings of the 28th International Conference on Very Large Databases (VLDB'02)* (pp. 287-298).

Xiong, X., Mokbel, M. F., & Aref, W. G. (2005). SEACNN: Scalable processing of continuous k-nearest neighbor queries in spatiotemporal databases. In *Proceedings of International Conference on Data Engineering (ICDE)* (pp. 643-654).

Yu, X., Pu, K. Q., & Koudas, N. (2005). Monitoring k-nearest neighbor queries over moving objects. In *Proceedings of International Conference on Data Engineering (ICDE)* (pp. 631-642).

KEY TERMS

Continuous Query: Query that needs to be evaluated at every time instant in order to ensure the correctness and validity of the query answer

Incremental Evaluation: Query evaluation that determines the incremental result containing only the changes of the results from the last evaluation.

K-Nearest Neighbors Query: Spatial query whose condition is given as a k nearest neighbors to the specified spatial object.

Location-Based Services (LBS): Services that deliver geographic information and geo-processing services to the mobile/stationary users taking into account their current location and references, or locations of the stationary/mobile objects of their interests.

Moving Object: Spatial object that continuously changes its location and/or other spatial properties: shape, dimensions, orientation, and so forth.

Negative Update: The update to the previous query answer containing objects that need to be removed from the query answer.

Positive Update: The update to the previous query answer containing objects that need to be added to the query answer.

Range Query: Spatial query whose condition is given by a spatial range or a distance from the specified spatial object.

Multilingual SMS

Mohammad Shirali-Shahreza
Sharif University of Technology, Iran

INTRODUCTION

In 1985, Ernie made the first telephone call on the mobile phone in Britain. In less than two decades, however, the mobile phone has turned into a necessary device for people, and now one out of every six individuals throughout the world has a mobile phone.

With the expanding use of mobile phones and the development of mobile telecommunications, telecommunication companies as well as companies manufacturing mobile phones decided to add additional features to their telephone sets in order to attract more customers. One of the services that were provided on the mobile phone was the SMS.

The SMS (short message service) is the transfer and exchange of short text messages between mobile phones. The SMS is defined based on GSM digital mobile phones. According to the GSM03.40 standard (GSM, 2000), the length of the exchanged message is 160 characters at most which are saved in 140 bytes depending on how information is saved according to the standards. These messages may be a combination of digits and letters or saved in non-text binary form. Using the same binary messages, one can also send pictures as well. The pictures, however, are two color and have a low quality (Shirali, 2006).

SMS messages are exchanged indirectly and through a component known as the SMSC. SMS messages have the following advantages:

- Communication is possible when the network is busy.
- We can exchange SMS messages while making telephone calls.
- We can send offline SMS messages.
- Various services are provided such as e-commerce.

One can also receive reports on the status of the SMS message or define a validity period for the SMS message (Nokia, 2001).

SMS PICTURE MESSAGE

The size of the SMS picture message is 72×28 pixels and it is two color. The saving format of the SMS picture is OTA. The structure of this format is as follows (Nokia, 2001).

The header of this format containing four fixed bytes is as follows:

Byte 1) 0000 0000 (→ 0)

Byte 2) 0100 1000 (→ 72)

Byte 3) 0001 1100 (→ 28)

Byte 4) 0000 0001 (→ 1)

As you can see in the above header, the second and third bytes indicate the height and width of the picture.

The structure of the body of the picture contains the pixels in 0 and 1. The amount of each pixel is saved in one bit. In each bit, 0 indicates the black and 1 the white color. Thus, every 8 pixels are saved in one byte. The order of saving of the pixels is from the left to the right and from the top to the bottom of the picture. Considering the size of the picture, the entire size of an SMS picture message is 256 bytes (see Figure 1).

SENDING SMS IN LOCAL LANGUAGE

Using SMS is not limited to the subscribers inside a country, and all mobile phone owners in other countries can also receive SMS.

In the early days, mobile phones supported limited languages such as English, but gradually other languages were also added to the potentialities of mobile phones. Today, mobile phone producers offer support of local language of the country where the phone set is to be supplied. For example, the mobile phones supplied to the Iranian market support Persian and Arabic languages as well. Thus, it is possible to send SMS in local languages. Anybody can send SMS in his own language and not need to use English (Stuiver, 2006).

Figure 1. Size of an SMS picture message

Image Size: $((72\times28 \text{ bit}) \div 8) \text{ byte} + 4 \text{ byte} = 256 \text{ byte}$

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

However, as already states, the mobile phones only support the English language plus the language of the concerned country (and sometimes a limited number of other languages). Thus, a person who lives outside his native country where the language spoken is different from his native language is not able to send or receive SMS in his mother tongue. In the same manner, those who live in a given country cannot send SMS in their local language to their friends in other countries, because the mobile phones of their friends may not support their local language, and therefore they have to send SMS in English. In some cases, people send their messages in their local language typed in English letters. For example, "place" is the English word for Arabic «مکان» "Makan." An Arab may type "Makan" (a transcription of the Arabic word in English characters) and send as SMS. This forces undesirable changes to the local languages and may, in the long run, destroy the language.

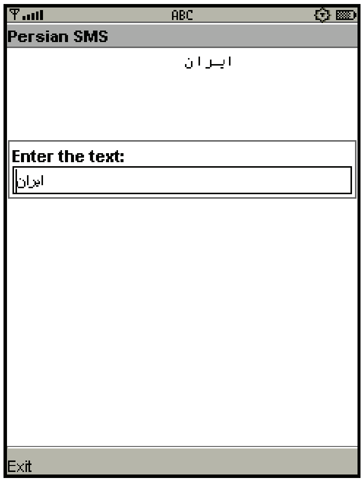
MY PROPOSED IDEA

In this article, I am presenting a solution to this problem. As indicated earlier, two-color pictures can be sent by SMS. Therefore, SMS picture messages can be sent to convey SMS in local languages. Thus, there will be no problem in view of support of the local language by the mobile phone of the person who receives the message. The details and procedure of this method are as follows.

First, the input text is received from the user. As the size of SMS is limited to 72 × 28 pixels, a limited amount of text can be incorporated into the picture. This amount differs for different languages. As the size of letters in Persian and Arabic is not identical, a precise size cannot be defined, but on average, two lines and in each line 9 Persian or Arabic letters can be accommodated. That is, a total of 18 Persian or Arabic characters can be placed in each picture. In case the mobile phone of the sender did not have the capability of entering text in the local language, a program can be developed to receive the texts in a local language. The possibility of supporting local language can be also added by installing some programs available in the market. For example, the Arabic/Farsi/Urdu Localization program produced by Psiloc Mobile Solution Company (Psiloc, 2006) can be used for Persian and Arabic languages. Of course, the text can be received from the user in the form of English characters and then converted into the local language, as done by the ArabTex program (Lagally, 2006). In this case, there is no need for a mobile phone capable of supporting the local language. However, this method is not recommended due to the defects previously indicated.

After receiving the text from the user, a certain number of letters depending on the language in use (due to limited size of the picture) are separated and saved in the picture of an SMS. This action is repeated until the entire text is saved in SMS picture messages. Then, these pictures are sent.

Figure 2. The program for converting text to picture SMS on the mobile phone



Persian SMS

ایران

Enter the text:

ایران

This idea has been implemented by J2ME (Java 2 Platform, Micro Edition). This language is a version of the Java programming language specially developed for small devices such as pocket PC computers, PDAs, mobile phones, and so on. As mobile phones extensively support the Java language, this method can be executed on a large range of mobile phones. It can also be executed on PDAs. A sample of this program for Persian and Arabic languages is shown in Figure 2.

ADVANTAGES

All mobile phones, even the black-and-white models and the old models, are capable of displaying picture SMS. Therefore, this method can be used without any hesitation. The cost of sending SMS is very low and sending a message in the form of a few SMS picture messages is not costly. This method is not only useful for sending SMS to another country. Even inside a given country, some old mobile phones do not support local language, and this method can solve their problem.

My proposed method has high security. For identification and tracing messages, an optical character recognition (OCR) program is needed to extract the text of the message and then to study the text. This process needs too many calculations and, considering the large volume of SMS exchanged daily, it is a difficult and time-consuming work.

In this method, special fonts of each language which may not be installed on the phone set of the receiver can be used to add to the beauty of the text.

REFERENCES

GSM. (2000). *GSM 03.40 v7.4.0. Digital cellular telecommunications system (phase 2+), technical realization of the Short Message Service (SMS) ETSI 2000.* Retrieved from http://www.etsi.org.

Lagally, K. (2006). *ArabTex: Multilingual computer typesetting.* Retrieved April 5, 2006, from http://www.informatik.uni-stuttgart.de/ifi/bs/research/arab_e.html

Nokia. (2001, May). *Sending content over SMS to Nokia phones.* Retrieved from http://www.forum.nokia.com

Psiloc. (2006). Retrieved April 5, 2006, from http://www.psiloc.com/

Shirali-Shahreza, M. (2006, April 11-13). Stealth steganography in SMS. *Proceedings of the 3rd IEEE and IFIP International Conference on Wireless and Optical Communications Networks (WOCN 2006)*, Bangalore, India.

Stuiver, M. (2006). *Sending SMS in foreign languages for example Arabic, Greek, Hebrew etc.* Retrieved April 5, 2006, from http://www.smswarehouse.com

KEY TERMS

Java 2 Platform, Micro Edition (J2ME): A collection of Java APIs targeting embedded consumer products such as PDAs, cell phones, and other consumer appliances.

Optical Character Recognition (OCR): Software that converts text scanned as a graphic into text a word processing program can use.

Short Message Service (SMS): A service for sending messages of up to 160 characters (224 characters if using a 5-bit mode) to mobile phones that use global system for mobile (GSM) communication.

Multimedia Contents for Mobile Entertainment

Hong Yan
City University of Hong Kong, Hong Kong
University of Sydney, Australia

Lara Wang
Tongji University, China

Yang Ye
Tongji University, China

INTRODUCTION

Electronic mobile devices are becoming more and more powerful in terms of memory size, computational speed, and color display quality. These devices can now perform many multimedia functions, including rendering text, sound, images, color graphics, video, and animation. They can provide users with great entertainment values, which were only possible with more expensive and bulky equipment before.

Technological advances in computer and telecommunications networks have also made mobile devices more useful for information exchange among users. As a result, a number of new mobile products and services, such as multimedia messages services (MMSs) and online games, can be offered to users by the industry.

It is commonly believed that "contents are the king" for multimedia products and services. As the mobile handsets and networks become more and more advanced, there is a stronger and stronger demand for high-quality multimedia contents to be used in new hardware systems. Content creation involves both computer technology and artistic creativity. Due to a large number of users, mobile entertainment has become an important part of the so-called creative industry that is booming in many countries.

In this article, we provide an overview of multimedia contents for mobile entertainment applications. The objective is for the readers to become familiar with basic multimedia technology, and the concepts and commonly used terms. Our focus will be on multimedia signal representation, processing, and standards.

SOUND

The original sound, such as speech from humans, can be represented as a continuous function $s(t)$. The sound can be recorded using electronic devices and stored on magnetic tapes. It can also be transmitted through telecommunications systems. The traditional telephone is used to send and receive waveform information of sound signals. The function $s(t)$ here is called an analog audio signal. This signal can be sampled every T seconds, or $f_s = 1/T$ times per second. The output signal is called a discrete-time signal. Each data sample in a discrete-time signal can have an arbitrary value. The sample values can be quantized so that we can represent them using a limited number of bits in the computer. These two processes, sampling and quantization, are called digitization, which can be achieved using an analog-to-digital converter (ADC) (Chapman & Chapman, 2004; Gonzalez & Woods, 2002; Mandal, 2003; Ohm, 2004). A signal $s(n)$ that is discrete both in time and in amplitude is called a digital signal. To render a digital sound signal, we must use a digital-to-analog converter (DAC) and send the analog signal to an electronic speaker.

The parameter $f_s = 1/T$ is called the sampling frequency. For telephone applications, usually $f_s = 8000$Hz, and for audio CD usually $f_s = 44100$Hz. To achieve stereo effects, the audio CD has two channels of data. There is only one channel in telephone applications. The sampling frequency f_s must be greater or equal to twice the signal bandwidth in order to reconstruct or recover the analog signal correctly. This is called the Nyquist sampling criterion. In telephone systems, a sound signal may have to be filtered to remove high-frequency components so that the sampling criterion is satisfied. This is why audio CDs have a higher sound quality than telephones.

The sound information can be stored as raw digital data. Some sound files, such as WAVE files (with extension ".wav") used on PCs, store raw sound data. This kind of format requires large storage space but has the advantage, since there is no information loss and the data can be accessed easily and quickly. To reduce the amount of data for storage and transmission, sound data are often compressed. Commonly used sound data compression methods include the m-law transformation, adaptive differential pulse code modulation (ADPCM), and Moving Picture Experts Group (MPEG) audio compression (Chapman & Chapman, 2004; Mandal, 2003; MPEG, n.d.; Ohm 2004).

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Currently, MPEG Audio Layer 3 (MP3) is a very popular sound data compression technique for mobile devices. MP3 is also a well-known sound file format. During data compression in MP3, a sound signal is decomposed into 32 frequency bands, and psychoacoustic models are used to determine the masking level and bit allocation pattern for each band. Modified discrete cosine transform (MDCT) is used to compress the data. The discrete cosine transform (DCT) has the so-called energy compact property—that is, it is able to pack most of the energy of a signal in a small number of DCT coefficients. For example, if $s(0) = 2$, $s(1) = 5$, $s(2) = 7$, and $s(3) = 6$, then the DCT coefficients are $S(0) = 10$, $S(1) = 3.15$, $S(2) = 2$, and $S(3) = 0.22$ (Mandal, 2003). In this case, from $S(0)$ to $S(3)$, the coefficients become smaller and smaller. We can simply retain the low-frequency components $S(0)$ and $S(1)$ and discard high-frequency ones $S(2)$ and $S(3)$ to obtain an approximation of the original signal. In practice, we can allocate more bits to code $S(0)$ and less and less bits for $S(1)$ to $S(3)$ to achieve a high compression ratio and at the same time maintain good signal quality.

Short sound files can be completely downloaded to mobile devices before being played. To reduce waiting time for downloading long sound files, audio streaming technology can be used (Austerberry 2005). In a streaming system, audio data is transmitted through a network and played by a mobile device as the data become available. That is, a sound does not have to be completely stored in the mobile device before being played. Steaming is useful if a large amount of data need to be received by a mobile device or live broadcasting is required.

We have focused on how to process sound waveform information above. In fact, sound can also be generated according to its parameters or a set of instructions. The musical instrument digital interface (MIDI) standard is used for such purpose (Chapman & Chapman, 2004; Mandal, 2003). A MIDI file contains information on what kind of instruments, such as different types of pianos, should be used and how they should be played. Several instruments can be arranged in different channels and played at the same time. MIDI files are much smaller than waveform-based sound files for music and is widely used for ring tones on mobile phones.

IMAGES

The imaging ability of mobile devices has been improved rapidly in recent years. Now most new mobile phones are equipped with digital cameras and can take pictures with millions of pixels. Software programs are available to edit an image, such as to enhance its contrast and change its color appearance. Mobile devices can also be used to send or receive images and browse the Web.

An important parameter for images is the resolution, which is closely related but should not be confused with the image size. Image resolution is usually measured by dots per inch (dpi). Higher resolution for the same physical area of an object would generate a larger image than lower resolution, but a large image does not necessarily mean a high resolution as it depends on the area that the image covers. The concept of resolution is often used in image printing and scanning. Typically, laser printers have a resolution of 300dpi or 600dpi and fax documents have resolutions from 100dpi to 300dpi.

A digital image can be considered as a two-dimensional (2D) discrete function $i(x, y)$. Like sound data, images need to be compressed to save storage space and transmission time. There are a number of methods that can be used to compress an image. Most methods are designed to reduce the spatial redundancy so that an image can be represented using a smaller amount of data. The most popular technique used for image compression is the Joint Photographic Experts Group (JPEG) standard (Gonzalez & Woods 2002). In JPEG-based compression, an image is divided into small, square blocks, and each block is transformed using the two-dimensional DCT (2D-DCT). Similar to audio data compression, the energy of a smooth image is concentrated in low-frequency components of the 2D-DCT coefficients. By allocating more bits to a small number of low-frequency components than to a large number of high-frequency components, we can achieve effective data compression.

In the JPEG2000 standard, the 2D discrete wavelet transform (2D-DWT) is used for image compression. In this method, an image is decomposed into several sub-bands, and different quantization schemes are used for different sub-bands. The 2D-DWT usually performs better—that is, it can provide a higher quality for similar compression ratio or a higher compression ratio for similar image quality than the 2D-DCT.

The JPEG and JPEG200 standards are usually used to provide lossy compression with a high compression ratio, although they can also be used for lossless compression. In lossy compression, the decompressed image is only an approximation of the original one, and as a result the image may appear to be blocky and blurred. The quality of an image from lossy compression can be improved using a number of techniques (Liew & Yan, 2004; Weerasingher, Liew, & Yan, 2002; Zou & Yan, 2005). In lossless compression, we can reconstruct or recover the original image exactly. Commonly used lossless compression methods include graphic interchange format (GIF), tagged image file format (TIFF), and portable network graphics (PNG). A GIF image can only show 256 colors and is especially useful for logos, buttons, borders, and simple animation on Web pages. GIF is a patented technology. In TIFF, image information is associated with different tags, which can be defined by users. TIFF is widely used for scanned office documents. PNG, which can support true colors, has been developed to be a royalty-free alternative to GIF. Similar to GIF, PNG can provide background transparency, but PNG does not support animation.

GRAPHICS

An image can also be generated from a set of drawing operations, similar to the way music is generated from MIDI instructions. Images represented by pixel values are called bitmaps, while those obtained from drawings are called vector graphics. The bitmap format is better for natural objects, such as human faces and outside sceneries, whereas the vector format is better for computer-generated objects, such as line drawings and industrial designs.

On computers and mobile devices, different operating or window systems provide different graphics utilities as software development tools. In general, they should all have graphics functions for drawing line shapes, such as lines, polygons, and circles, and displaying bitmaps. Some systems provide more advanced functions, such as geometric shape transformations, color gradients, and spline curve and surface displays. Currently, there are many proprietary mobile operating systems developed by different manufacturers. This makes it difficult to port graphics applications from one phone model to another model. However, a few systems have been adopted by more and more manufacturers. They include Symbian, Microsoft Windows Mobile OS, and Linux-based mobile OS.

There are also several graphics formats that are supported by many mobile operating systems. They include proprietary formats Java 2 Micro Edition (J2ME) and Flash, and royalty-free and open standard scalable vector graphics (SVG). J2ME is a general purpose programming language for small electronic devices and contains many useful graphics functions (Edward, 2003; Wells, 2004). It has gained widespread use in mobile phones, especially for games.

Flash was developed by Macromedia, which is now part of Adobe Systems. In addition to graphics, Flash can also be used to present other types of multimedia data, such as text, sound, and animation. The output file of Flash is called an SWF movie, which contains definition tags and control tags. The definition tags describe the objects in a movie, such as text, shapes, bitmaps, sound, and sprites. The control tags define how and when an object should be transformed and displayed. SWF movie files use a very compact binary format, so they can be transmitted over the Internet quickly. They have found many applications to Web page design and are now used more and more in mobile phones. Macromedia Flash provides an authoring tool to design SWF movies. One can also create a SWF movie using a computer program according to the SWF file format.

SVG is an open graphics standard developed by the World Wide Web Consortium (W3C). It is royalty free and vendor independent. SVG uses text format and provides a language to define graphics display, in a way similar to the HyperText Markup Language (HTML) that defines text display. SVG can be used to describe vector shapes, text, bitmap images, and animation. Since vector graphics can be easily scaled larger or smaller, SVG files can be used for different resolutions, such as in printing which requires a high resolution, and in displays on mobile devices which have low resolution. A number of graphics software packages, such as CorelDraw and Adobe Illustrator, can output SVG files.

VIDEO

A digital video contains a sequence of images $v(x, y, n)$, where x and y are spatial coordinates and n represents time. For each n value or a time sample, we have a frame of video or an image. In video compression, we can explore both spatial or intra-frame redundancy and temporal or inter-frame redundancy. Spatially, a frame can be compressed just like an image. Temporally, a frame is similar to its proceeding frame, so we can expect a higher compression ratio for video than for each image separately and independently.

A straightforward way of reducing the temporal redundancy is to use the proceeding frame and approximation of the current frame. A better approximation can be achieved if we take into account movements of the objects in the image sequence. We can divide an image into small blocks and search in the proceeding frame for a closest shifted version of each block. Then, we only need to code the difference between the block under consideration and its closest match in the previous frame, and the shift between the two blocks. This procedure is called motion estimation.

Extensive research has been carried out in the field of video processing on how to estimate and compensate motions efficiently.

A series of MPEG standards have been developed for video compression (MPEG, n.d.; Watkinson, 2004). MPEG-1 was developed for coding of moving pictures and associated audio for digital storage media, such as video CD and MP3, at up to 1.5 Mbit/s. It uses block motion compensation and the DCT to code the residual image. MPEG-2 provides generic coding of moving pictures and associated audio information for bit rates from 1.5 to 80 Mbit/s. It is an extension of MPEG-1, allows combinations of video and audio streams, and supports different packet formats for data transmission. MPEG-2 products include digital TV setup boxes and DVDs. MPEG-4 provides standardized technology for video and audio data storage, transmission, and content access and manipulation on digital TV, interactive graphics applications, and the World Wide Web. In addition to video and audio coding, it supports creation of synthetic objects. For example, mesh models can be used for human face animation. MPEG-7 is a standard for describing multimedia content, including images, graphics, 3D models, audio, speech, video, and the way various data should be combined in a multimedia presentation. MPEG-21, which is still under development, defines the multimedia framework for different users, including content creators, producers, distributors, and service providers, to access, exchange, manipulate, and trade a large variety of multimedia items.

The International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) has developed a series of standards H.26x for video phone and videoconference applications. H.261 supports data rates as multiples of 64Kbit/s, that is, $p\times64$Kbit/s, where $1 \leq p \leq 30$. H.261 only supports two image frame sizes, common interchange format (CIF) or 352×288 pixels, and quarter common interchange format (QCIF) or 176×144 pixels. H.262 is the same as the video part of MPEG-2. H.263 provides a number of improvements over H.261, MPEG-1, and MPEG-2. For example, it uses better methods for motion compensation, offers higher video quality, and supports more image sizes. H.264 is the same as MPEG-4 Part 10, advanced video coding (AVC), and is jointly developed by the ITU-T Video Coding Experts Group (VCEG) and MPEG. H.264/MPEG-4 AVC employs a number of new techniques, such as variable block-size motion compensation (VBSMC), to improve the compression performance. It is also more flexible under a variety of network environments. H.263 and H.264 have been used in 3GP movies for second-generation (2G) and third-generation (3G) mobile phones.

Similar to audio, video can also be streamed (Austerberry, 2005). This is especially useful for video phone calling, videoconferencing, and live broadcasting. Current 3G mobile networks already offer video calls and many video-based entertainment programs, such as news and sports, which were only possible through TV before.

ANIMATION

Animation means presentation of a sequence of artificially created images. In video, the images are obtained from a camera, while in animation, the images are drawn by hand or generated by the computer. Moving pictures in animation are often synchronized with audio to create movies.

Cartoon movies had been used for entertainment long before the digital computer was invented. A cartoon film must contain 24 picture frames every second to show smooth motions (Chapman & Chapman, 2004). This means 86,400 pictures for a one-hour movie; obviously it is very labor intensive. The labor cost can be reduced using cel animation and key-frame animation techniques. In cel animation, the still background is drawn only once and the moving part is drawn frame by frame, each on a cel, a sheet of transparent material, which is placed on the background picture. In key-frame animation, a motion is decomposed into important key frames and in-betweens. Experienced animators are assigned to draw the key frames and junior animators the in-betweens. Now using the computer, the in-betweens can often be generated using pattern matching and interpolation techniques.

Simple animation can be displayed on the computer or mobile devices by simply going though a sequence of images. GIF provides such function. It is useful for small animated pictures, which does not require many colors or audio. It is widely used for Web page design and for MMS on mobile phones.

Simple animation can also be generated using morphing techniques. In Macromedia Flash, one shape can be morphed to another shape specified by a set of parameters. Each shape is described by its edges and color information. In this method, key frames are drawn according to shape specifications, and in-betweens are automatically generated by the computer through the morphing process.

Three-dimensional (3D) animation has been used in many digital entertainment products and services. In 3D animation, an object is often represented using a mesh model. It is shaded according to light and camera settings, and may be superimposed with a texture. Now there are a number of 3D design packages available for PCs as well as powerful workstations. Virtual reality (VR) systems make extensive use of 3D graphics and animation. The Virtual Reality Modeling Language (VRML) was developed to support VR on the Web (Ames, Nadeau, & Moreland, 1997).

An interesting and challenging task in computer animation is to automatically animate the human and animal faces (Huang & Yan, 2002, 2003; Parke & Waters, 1996; Yan, 2001). This involves synchronizing the mouth movement with voice and generating different facial expressions. We have recently developed a real-time lip synchronization and facial animation system, which has already been used by several Internet and communications companies for 3G and MMS applications on mobile phones (Tang, Liew, & Yan 2005). Our system can create many virtual characters with different styles. It can match voice signals with lip/mouth shapes smoothly and automatically. The lip-sync can be done in real time (e.g., 20 frames per second) for several languages, including Chinese (Cantonese and Mandarin), English, French, German, Japanese, and Spanish. In addition, a virtual character can express his/her emotion with different kinds of facial expressions. Examples of the movies produced by our system can be found on http://www.HyperAcademy.com or simply http://www.hy8.com.

TECHNOLOGY TRENDS

Mobile technology is rapidly advancing. We expect to see many new or enhanced mobile entertainment products and services in the next several years. First of all, future mobile devices will have higher computational power for its processors, higher resolution for digital cameras, and higher data rates through the telecommunications network. Currently, there is criticism for 3G mobile phones that the battery does not last long for displaying movies. To support multimedia functions for a longer period of usage, mobile devices need improved technologies for reduced power assumption and better batteries.

Traditionally, most multimedia contents based on graphics, video, and animation are produced for TV and other large display screens. Simple image down sampling of the images can cause visibility problems. So future multimedia editing and authoring tools should take into account the legibility of the material for mobile devices. An interesting research topic in image processing and pattern recognition is how to reduce the size of a bitmap image optimally so that the output is most legible. This requires sophisticated algorithms for feature detection and pattern extraction from the image.

More and more multimedia contents will become available for mobile devices in the future. Mobile gaming is one of the rapidly growing areas. There will also be increased use of mobile phones for music, movies, advertisements, news reports, storytelling, finance, weather and traffic information, and educational and training programs, and for access to the Web, most of which are similar to the capabilities and functions provided by TV and desktop computers. For Web applications, better user interface and page layout will be needed so that the users can find information needed easily. We also expect increased applications of video phone calls, videoconferencing, and video e-mail messages.

A more challenging task for a mobile device to perform is to recognize its user's voice, speech, face, and even facial expressions and emotions, and interact with the user. Considerable progress has been made in the past in these areas, but the recognition reliability still needs significant improvement. These problems can be solved to some extent under restricted conditions. For mobile phones, it is difficult to control the background noise or appearance for normal usages, so more robust pattern recognition techniques must be developed to solve these problems.

ACKNOWLEDGMENTS

This work is supported by a research grant from City University of Hong Kong (Project 9610034).

REFERENCES

Ames, A. L., Nadeau, D. R., & Moreland, J. L. (1997). *VRML 2.0 sourcebook.* New York: John Wiley & Sons.

Austerberry, D. (2005). *The technology of video and audio streaming.* Burlington, MA: Focal Press, Elsevier.

Chapman, N., & Chapman, J. (2004). *Digital multimedia.* Chichester, UK: John Wiley & Sons.

Edward, J. (2003). *J2ME: The complete reference.* New York: McGraw-Hill.

Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing.* Upper Saddle River, NJ: Prentice Hall.

Huang, D., & Yan, H. (2002). Modeling and animation of human expressions using NURBS curves based on facial anatomy. *Signal Processing: Image Communications, 17,* 457-465.

Huang, D., & Yan, H. (2003). NURBS curve controlled modelling for facial animation. *Computers & Graphics, 27,* 373-385.

Liew, A., & Yan, H. (2004). Blocking artifacts suppression in block-coded images using overcomplete wavelet representation. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(4), 450-461.

Mandal, M. K. (2003). *Multimedia signals and systems.* Boston: Kluwer Academic.

MPEG. (n.d.). *Homepage.* Retrieved from http://www.chiariglione.org/mpeg/

Ohm, J.-R. (2004). *Multimedia communication technology, representation, transmission and identification of multimedia signals.* New York: Springer.

Parke, F. I., & Waters, K. (1996). *Computer facial animation.* Wellesley, MA: A.K. Peters.

Tang, J. S. S., Liew, A., & Yan, H. (2005). Human face animation based on video analysis, with applications to mobile entertainment. *Journal of Mobile Multimedia, 1*(2), 132-147.

Watkinson, J. (2004). *The MPEG handbook: MPEG-1, MPEG-2, MPEG-4.* Burlington, MA: Focal Press, Elsevier.

Weerasinghe, C., Liew, A., & Yan, H. (2002). Artifact reduction in compressed images based on region homogeneity constraints using projection on to convex sets algorithm. *IEEE Transactions on Circuits and Systems for Video Technology, 12*(10), 891-897.

Wells, M. J. (2004). *J2ME game programming.* Boston: Thomson Course Technology.

Yan, H. (2001). Image analysis for digital media applications. *IEEE Computer Graphics and Applications, 21*(1), 18-26.

Zou, J. J., & Yan, H. (2005). A delocking method for BDCT compressed images based on adaptive projections. *IEEE*

Transactions on Circuits and Systems for Video Technology, 15(3), 430-435.

KEY TERMS

Audio Signal Processing: Acquisition, compression, enhancement, filtering, transformation, and transmission of sound data.

Computer Graphics: Modeling and rendering of two-dimensional or three-dimensional objects on the computer.

Digital Entertainment: Providing multimedia materials, such as music and movies, on digital devices, such as computers and mobile phones.

Facial Animation: Modeling and displaying talking human faces with different expressions.

Image Processing: Acquisition, compression, enhancement, filtering, transformation, and transmission of pictures.

Lip-Synchronization: Matching the mouth shape created by the computer with the voice signal.

Multimedia Content: Digital data of text, music, graphics, images, and video

Video Processing: Acquisition, compression, ehancement, filtering, transformation, and transmission of movies.

Multimodality in Mobile Applications and Services

Maria Chiara Caschera
Istituto di Ricerche sulla Popolazione e le Politiche Sociali – CNR, Italy

Fernando Ferri
Istituto di Ricerche sulla Popolazione e le Politiche Sociali – CNR, Italy

Patrizia Grifoni
Istituto di Ricerche sulla Popolazione e le Politiche Sociali – CNR, Italy

INTRODUCTION

Multimediality and multimodality are concepts with multiple meanings. In Van den Anker and Arnold (1997), multimediality is defined as a way to present and convey information using several different media. Multimodality provides the user with multiple modalities of interacting with a system, beyond the traditional keyboard and mouse.

Both multimediality and multimodality refer to more than one communication channel. The diffusion of mobile devices and the development of their services and applications is connected with the natural communication approach preferred by users, which combines several modalities (speech, sketch, etc.) in order to communicate. It is therefore necessary to integrate the diverse modalities so that the features of the mobile device are more similar to the paradigm of human communication, making it simpler to use.

W3C (World Wide Web Consortium) activities have solved various mobile Web problems affecting the diffusion of mobile devices, such as practice Web navigation from mobile devices, multimodal interaction for the mobile Web, and multimedia and graphics for multimedia messaging. Diverse W3C working groups are involved in the discussion about device independence, multimodal Web access, and type of content for multimedia messaging. Characteristics identified by W3C according to the power and extensibility of XML (eXtensible Markup Language) enable the exchange of rich multimedia content and promote inter-user communication.

The greater opportunity to access and exchange information according to the content and opportunities offered by mobile devices are the two main elements of the growing interest in added-value services in different social environments. In fact, the opportunities offered by multimedia and multimodal technologies are important for any type of communication services, and as these technologies allow new ways of communicating, particular attention can be devoted to people with no technological ability or those with disabilities. This article introduces and discusses the problems and future scenarios and prospects of multimodality and mobile communication, analyzing the multimodal dialogue systems.

BACKGROUND

Multimodal applications combine visual information (involving images, text, sketches, and so on) with voice, gestures, and other modalities to provide powerful mobile applications, giving users the flexibility to choose one or more of the multiple interaction modalities. These systems break down the barriers in adopting mobile devices for added-value services.

Mobile applications and services generally involve both multimodality and multimediality, using different modalities and (as specified below) different communication channels. Two typical cases of mobile multimedia use are person-to-person communication and person-to-content communication (Ericsson, 2004). In addition, several applications require multimedia data, and numerous studies have been devoted to developing new ideas for methodologies, technologies, and algorithms for indexing, retrieving, compressing, transmitting, and integrating different types of data (Pham & Wong, 2004).

The use of multimodality and multimediality in mobile devices allows a simple, intuitive communication approach and generates new, richer services for users (Colby, 2002). When developing multimodal services, it is essential to consider perceptual speech, audio, and video quality for optimum communication system design and effective transmission planning and management in order to satisfy customer requirements (Kitawaki, 2004).

The following parameters should be considered when characterizing quality in advanced mobile devices: (1) wideband speech, audio, and video for multimedia; (2) noise reduction; and (3) speech recognition-synthesis for hands-free communication.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Multimedia and multimodal systems use different channels of communication, and specific devices are developed to increase the information flow between user and system. Nigay and Coutaz (1993) distinguish between the two, observing that a multimodal system is able to automatically model information content through a high level of abstraction. This difference leads to the definition of two main characteristics of multimodal interfaces:

- fusion among different data types and different input/output devices; and
- temporal constraints, imposed by information processing to and from input/output devices.

A multimodal system is an hw/sw system that allows one to receive, to interpret, and to process input, and that generates as output two or more interactive modalities in an integrated and coordinated way.

Communication among people is often multimodal, and it is obtained combining different modalities. Multimodal interfaces allow several modalities of communication to be harmoniously integrated, making the system's communication characteristics more similar to the human approach.

Multimodal interfaces provide the user with multiple interaction paradigms through different types of communication input. Data fusion is one of the main problems in human-computer interaction, where each datum is generated through a distinct interaction mode. Furthermore, the management of these multiple processes includes synchronization and selection of the predominant mode. Consequently, an important issue in multimodal interaction is the integration and synchronization of several modalities in a single system. In literature two approaches are often used: signal fusion, and information fusion at the semantic level.

The first approach is preferred for matching and synchronizing modalities as speech and labial movement. The semantic fusion is used for modalities that differ in a temporal scale. In this approach, time is very important because chunks of information with different modalities are considered, and integrated if they are temporally close. The integration can be carried out using an intermediate approach between the signal integration and the semantic fusion.

The relation among modal components can be classified as follows (Bellik, 2001):

- **Active:** (Act in following tables)—when two events, produced by two different devices, cannot be completely and correctly interpreted without ambiguities if one of the two events is unknown.
- **Passive:** (Pas in following tables)—when an event produced by a given device cannot be completely and correctly interpreted without ambiguities if the state of the other devices is unknown.

The input synchronization of a multimodal system can be defined as:

- **Sequential:** (Seq in following tables)—if the interpretation of the interactive step depends on one mode and the modalities can be considered one by one.
- **Time-Independent Synchronized:** (TIS in following tables)—if the interpretation of the interactive step depends on two or more modalities and the modes are simultaneous.
- **Time-Dependent Synchronized:** (TDS in following tables)—if the interpretation of the interactive step depends on two or more modalities and the semantic dependence of the modalities has a close temporal relationship.

There are several levels of synchronization (W3C):

- **Event-Level:** If the inputs of one mode are received as events and immediately propagated to another mode.
- **Field-Level:** If the inputs of one mode are propagated to another mode after a user has changed the input field or the interaction with a field is terminated.
- **Form-Level:** If the inputs of one mode are propagated to another mode after a particular point of the interaction has been achieved.
- **Session-Level:** If the inputs of one mode are propagated to another mode after an explicit changeover of mode.

The semantic fusion of modal input occurs in two steps: (1) the first matches the modalities to obtain a low-level interpretation module, by grouping the input events in multimodal events; and (2) the second transfers the multimodal inputs to the high-level interpretation module, in order to obtain the meaning of their events. This high-level interpretation defines the type of actions that will be triggered by the user and the used parameters. These parameterized actions are passed to the application dialog manager to start their execution.

MAIN FOCUS OF THE ARTICLE

There is an emerging need for integration among the various input modalities, through signal integration and semantic fusion, and an additional need to disambiguate the various input modalities and coordinate output modalities, to enable the user to have a range of integrated, coordinated interaction modalities.

Martin and Toward (1997) proposed a theoretical framework for studying and designing multimodal systems based on a classification of six basic types of cooperation between modalities:

Table 1. Correspondences among types of cooperation, relation among modal components, and input synchronization

Types of Cooperation	Relational Among Modal Components	Input Synchronization
C	Act, Pas	TDS, TIS
E	No relation	Seq
R	Act	TDS, TIS
T	Pas	Seq
CC	No relation	No synchronization
S	No relation	Seq

- **Complementarity:** (C in following tables)—different chunks of information comprising the same command are transmitted over more than one mode.
- **Equivalence:** (E in following tables)—a chunk of information may be transmitted using more than one mode.
- **Redundancy:** (R in following tables)—the same chunk of information is transmitted using more than one mode.
- **Transfer:** (T in following tables)—a chunk of information produced by one mode is analyzed by another mode.
- **Concurrency:** (CC in following tables)—independent chunks of information are transmitted using different modalities and overlap in time.
- **Specialization:** (S in following tables)—a specific chunk of information is always transmitted using the same mode.

Benoit, Martin, Pelachaud, Schomaker, and Suhm (2000) have advanced several proposals for integrating different modalities. Table 1 presents correspondences among types of cooperation, input synchronization, and relation among modal components.

An approach for classifying multimodal applications considers the services provided. Aslan, Xu, Uszkoreit, Krüger, and Steffen (2005) classify services into three groups: information services, transactional services, and composed services. Information services can provide the user with information such as air travel info, city info, and so on. Transactional services, such as travel planning services, involve a much closer engagement and a greater degree of two-way interactivity. Composed services integrate various services and information for complex situations. Starting from the multimodal applications taxonomy introduced in Benoit et al. (2000) based on several proposals for multimodal systems in different application domains, Table 2 provides a classification of services by input mode, domain of application, cooperation between modalities, type of multimodal fusion, input synchronization, and relation among modal components. It considers the following input modalities (Benoit et al., 2000): automatic speech recognition (ASR), gesture recognition (GR), hand-writing recognition (HR), pointing (P), eye tracker (ET), gaze tracking (GT), keyboard (K), 2D or 3D controller device (2D and 3D), object recognition (OR), speaker verification (SV), and face recognition (FR).

The application domains more frequently involved in the information services include road maps, geographical maps, tourist information, and so on. These often involve the integration of automatic speech recognition and gesture recognition—that is, these two modalities can be fused at a semantic level.

Transactional services often involve speech, gesture, and handwriting recognition; cooperation between these modalities is complementary and equivalent. Multimodality provides end users with the option to move among different modalities. They might send a message by voice or text, using a mobile phone or personal digital assistant (PDA). Users can dynamically select the most appropriate interaction mode for their needs by multimodal services, improving service accessibility and usability, especially for people with disabilities.

Table 2 highlights the fact that mode fusion can take place at the semantic level, intermediate level, or signal.

The synchronization level depends on the temporal granularity of the multimodal interaction. Diverse modalities can be combined, using several levels of synchronization according to the expected temporal granularity. Table 2 therefore does not include the level of granularity synchronization.

In literature there are several approaches to represent the events of multimodal inputs:

- **Typed Feature Structures** (Cohen et al., 1997; Johnston et al., 1997)**:** In this approach, multimodal inputs can be transformed into typed feature structures that represent the semantics attributed to the various modalities. These feature structures are combined by unification (QuickSet).
- **Syntactic Representation** (Faure & Julia, 1993)**:** The input multimodal events are represented as a triplet {verb, object, location}. This representation is sufficient for input in the form of speech with deictic references, but it is not clear how this approach can be extended to deal with several modal events.
- **Melting Pots** (Nigay & Coutaz, 1995)**:** A melting pot encapsulates types of structural parts of a multimodal event. The content of a structural part is a time-stamped piece of information. The melting-pots are constructed by events in elementary inputs with different mechanisms of fusion: micro-temporal, macro-temporal, and contextual fusion.

Table 2. Classification of services by domain of application, input modalities, cooperation between modalities, type of multimodal fusion, relation among modal components, and input synchronization

	Application/Domain	Input	Cooperation	Fusion	Relation among Modal Components	Input Synchronization
Information Services	Air Travel Info	ASR, P, K	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Web Navigation	ASR, P, K	E	n/a		Seq
	Geographic Maps and Blocks World	ASR, 2D-GR, 3D-GR, GT	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Tourist Information	ASR, 2D-GR and P, HR	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Office Assistance	ASR, 2D-GR and P, HR	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Tourist Map	ASR, P, K	all forms	semantic	Act, Pas	Seq, TDS, TIS
	Consumer Information	ASR, K	n/a	n/a		
	Product Information	ASR, K, GR	E, R, C, S	semantic	Act, Pas	TDS, TIS
	Information Retrieval	ASR, 2D-GR, K	C	semantic	Act, Pas	TDS, TIS
	Campus Information	ASR, GR	E, C, S	semantic	Act, Pas	Seq, TDS, TIS
	Notebook	ASR, P, K	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Talking Agent	ASR, GR	n/a	signal intermediate semantic		
	Train Scheduling	ASR, GT	R	intermediate	Act, Pas	TDS, TIS
	Road Maps	ASR, P	E, S	semantic		Seq
Transactional Services	Notebook	ASR, P, K	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Videoconferences	ASR, FR	E, R, C, S	semantic	Act, Pas	Seq, TDS, TIS
	Train Scheduling	ASR, GT	R	intermediate	Act	TDS, TIS
Composed Services	Interactive TV	ASR, 3D-GR	S	n/a		
	Robot Control	ASR, 2D-GR and P, HR	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Image Analysis	ASR, 2D GR and P, HR	C, E	semantic	Act, Pas	Seq, TDS, TIS
	Emergency Dispatch	ASR, 2D GR and P, HR	C, E	semantic	Act, Pas	Seq, TDS, TIS

- **Micro-Temporal Fusion:** Combines two information units produced concurrently or very close to one another. *Macro-Temporal Fusion* combines sequential or temporally close information units, when these units are complementary. *Contextual Fusion* combines information units according to semantic constraints.
- **Partial Action Frame** (Vo & Wood, 1996; Vo & Waibel, 1997; Vo, 1998)**:** The input of each mode is separately interpreted and then analyzed and transformed into a semantic frame containing slots that specify the control parameters. Information in the partial action frames can be incomplete or ambiguous. Each sequence of grouped input events has a score based on their mutual information.

Semantic fusion is not further defined by the characteristics of the events because the above approaches can be used indifferently. For example, speech and pen fusion can be achieved by both Typed Feature Structures and Partial Action Frame. In literature, events for fusing gesture and speech have been represented by Typed Feature Structures or Syntactic Representation. The Melting Pots approach can be used to fuse speech, keyboard, and mouse.

The diffusion and implementation of multimodal services is supported by the activities of the World Wide Web Consortium (W3C), aimed at extending the modalities of interaction for mobile devices and particularly devoted to solving various mobile Web problems:

- minimization of costs for mobile devices,
- practice Web navigation from mobile devices,
- multimodal interaction for the mobile Web, and
- multimedia and graphics for multimedia messaging.

Some W3C working groups focus their activities on issues such as independence from devices, multimodal Web access, and types of content for multimodal messaging. These specifications allow rich multimodal contents to be transmitted and are based on the power and extensibility of XML (eXtensible Markup Language).

W3C has defined specific modal languages for developing Web-based applications. The W3C groups voice browser and multimodal interaction are working to standardize these languages. VoiceXML and speech application language tags (SALT) are valid instruments to support the implementation of visual-representation-enriched systems and visual browsers, with instruments such as eXtensible Hypertext Markup Language (XHTML), Cascading Style Sheet (CSS), Synchronized Multimedia Integration Language (SMIL), and scalable vector graphics (SVG). In addition, new trends are outlined to allow intermodal interaction by instruments as Ink Markup Language (InkML), which captures pen movements and enables data exchange by "digital ink." Another interesting language is extensible multimodal markup annotations (EMMA), used to implement semantic interpretations of a great variety of inputs, such as voice, text languages, and digital ink. The Speech Services Control (SpeechSC) working group of the Internet Engineering Task Force (IETF) develops protocols to support distributed speech recognition, speech synthesis, and speaker verification services, and expects to take advantage of W3C's work on speech recognition grammar specification (SRGS), Speech Synthesis Markup Language (SSML), and semantic interpretation (SI).

FUTURE TRENDS

The trend toward converging various methodologies and technologies has created new mobile devices, which make available complex services and contribute to the sharing of experiences and the inclusion of people as members of the community. The usability, accessibility, portability, wearability, and power of future mobile devices thus give rise to important perspectives for the evolution of services in different application domains.

When considering future trends, it is important to consider the developments offered by methodologies and technologies on the semantic fusion of modal inputs. There are several approaches for representing events of modal inputs: melting pots, partial action frame, syntactic representation, and typed feature structures. Other approaches may be defined for future multimodal systems adapted to the next generation of mobile devices. Research into the integration of different interaction modalities is of great interest for the diffusion and connectivity of mobile devices.

CONCLUSION

The development of multimodal tools is producing notable interest in mobile devices, especially for accessing information and performing transactions. This article discussed the methods and future prospects of mobile multimodal applications and services. These are divided into three groups: information services, transaction services, and composed services, which are then further classified into some important characteristics, such as application domain, input and output modalities, type of cooperation between modalities, and type of multimodal fusion.

Finally, special attention is dedicated to the activities of the World Wide Web Consortium (W3C) devoted to extending multimodal services on the Web and on mobile devices.

REFERENCES

Aslan, I., Xu, F., Uszkoreit, H., Krüger, A., & Steffen, J. (2005). COMPASS2008: Multimodal, multilingual and crosslingual interaction for mobile tourist guide applications. *Proceedings of INTETAIN—Intelligent Technologies for Interactive Entertainment* (pp. 3-12). Madonna di Campiglio, Italy.

Bellik Y. (2001). Technical requirements for a successful multimodal interaction. *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue* (online)*,* Verona, Italy.

Benoit, C., Martin, J. C., Pelachaud, C., Schomaker, L., & Suhm, B. (2000). Audio-visual and multimodal speech systems. In D. Gibbon (Ed.), *Handbook of standards and resources for spoken language systems: Supplement volume* (pp. 1-95). NY: Walter De Gruyter Inc.

Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., & Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications. *Proceedings of the 5th ACM International Multimedia Conference* (pp. 31-40). Boston: ACM Press/Addison-Wesley.

Colby, J. (2002). *Multimodality: The next wave of mobile interaction.* Retrieved from http://www.comverse.com

Faure, C., & Julia, L. (1993). Interaction homme-machine par la parole et le geste pour l'édition de documents: TAPAGE. *Proceedings of the International Conference on Interfaces to Real and Virtual Worlds* (pp. 171-180).

Ericsson. (2004). *Mobile multimedia: The next step in richer communication.* Retrieved from http://www.ericsson.com

Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. H., Pittman, J. A., & Smith, I. (1997). Unification-based multimodal integration. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (pp. 281-288). Madrid, Spain.

Kitawaki, N. (2004). Perspectives on multimedia quality prediction methodologies for advanced mobile and IP-based telephone. *Proceedings of the Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction* (pp. 1-8). Mainz, Germany.

Martin, J. C. (1997). Toward intelligent cooperation between modalities: The example of a system enabling multimodal interaction with a map. *Proceedings of the International Joint Conference on Artificial Intelligence Workshop on Intelligent Multimodal Systems* (online)*,* Nagoya, Japan.

Nigay, L., & Coutaz, J. (1993). A design space for multimodal systems—Concurrent processing and data fusion. *Proceedings of INTERCHI'93—Conference on Human Factors in Computing Systems* (pp. 172-178). Boston: Addison-Wesley.

Nigay, L., & Coutaz, J. (1995). A generic platform for addressing the multimodal challenge. *Proceedings of the International Conference on Computer-Human Interaction* (pp. 98-105). Boston: ACM Press.

Pham, B., & Wong, O. (2004). Handheld devices for applications using dynamic multimedia data. *Proceedings of Graphite* (pp. 123-130). Singapore.

Van den Anker, F. W. G., & Arnold, A. G. (1997). Mobile multimedia communication: A task- and user-centered approach to future systems development. *Proceedings of the 7th International Conference on Human Computer Interaction* (pp. 651-654).

Vo, M. T. (1998). *A framework and toolkit for the construction of multimodal learning interfaces.* PhD Thesis, Carnegie Mellon University, USA.

Vo, M. T., & Waibel, A. (1997). *Modeling and interpreting multimodal inputs: A semantic integration approach.* Technical Report CMU-CS-97-192, Carnegie Mellon University, USA.

Vo, M. T., & Wood, C. (1996). Building an application framework for speech and pen input integration in multimodal learning interfaces. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 3545-3548).

W3C. (2003, January 8). *Multimodal interaction requirements.* Retrieved from http://www.w3.org/TR/2003/NOTE-mmi-reqs-20030108/

KEY TERMS

Extensible Multimodal Annotation (EMMA): This markup language is a W3C recommendation based on XML used for information automatically extracted from the user input. This language is focused on annotating interpretation of information in input.

InkHTML: A markup language devoted to describing ink data acquired with an electronic pen.

Internet Engineering Task Force (IETF): Activity devoted to promote and develop Internet standards.

Mobile Device: A pocked-sized computer device used for managing and accessing information and communicating in the different context of life.

Personal Digital Assistant (PDA): A mobile device used as a personal organizer, a data book, a task list, a clock, and a calculator.

Scalable Vector Graphics (SVG): A W3C Markup Language describing a two-dimensional vector graphic. It allow shapes, raster graphics images/digital images, and text.

Speech Application Language Tag (SALT): A markup language used for adding voice recognition to HTML and XHTML files.

Speech Synthesis Markup Language (SSML): A W3C recommendation based on XML used for speech synthesis.

Synchronized Multimedia Integration Language (SMIL): A W3C recommendation based on XML used for multimedia presentation.

VoiceXML: The W3C standard XML format used for specifying voice dialogues between a human and a computer.

Multi-User OFDM in Mobile Multimedia Network

Ibrahim Al Kattan
American University of Sharjah, UAE

Habeebur Rahman Maricar
American University of Sharjah, UAE

INTRODUCTION

The optimization of resource allocation in multi-user orthogonal frequency division multiplexing (OFDM) has been of great concern to the mobile commerce. The major concern is the allocation of subcarrier and power to different users in order to minimize the total transmitted power, consequently increasing the total data rate and improving the performance of the wireless communication system. Multi-user OFDM systems have been in use recently and seem promising for future mobile network applications. This research would mainly focus on modeling and improving the wireless communication system with multi-user OFDM. The proposed optimization techniques for OFDM wireless communication will be conducted in two phases. Phase one is to optimize the subcarrier allocation to mobile hosts, and the second phase is to optimize the power allocation (depends on the number of transmitted bits) in a multi-user OFDM system.

A wireless and mobile multimedia network has become challenging in the cutting-edge technology of today's global market economy. Some current applications include mobile computing, mobile phones, satellite communications, radio stations, and so forth. An efficient communication plays a very important role in any wireless and mobile network. Global system for mobile communications (GSM) and code-division multiple access (CDMA), commonly known as the second-generation (2G) mobile systems, are being widely used all over the world and have been constantly developing over the last decade (Zheng, Huang, & Wang, 2005). These developments have led to the growth of third-generation (3G) and fourth-generation (4G) mobile systems. Among the many technologies proposed for 4G systems, orthogonal frequency-division multiplexing (OFDM) has been of great interest over the past decade (Gross, Geerdes, Karl, & Wolisz, 2005). The major reason is that it can provide high data rates over a wireless channel and it divides the bandwidth into subcarriers (Gross et al., 2004).

OFDM is defined as a multi-carrier transmission technique that has been recognized as an excellent method for high-speed bi-directional wireless data communication. OFDM effectively squeezes multiple modulated carriers tightly together, thus reducing the required bandwidth. The modulated signals overlap with each other, but they do not interfere with each other since they are kept orthogonal (Intel in Communications, n.d.). Figure 1 shows a conventional frequency division multiplexing (FDM) with nine subcarriers using filters, while Figure 2 shows an OFDM with nine subcarriers. The quality remains the same, but the bandwidth required and consequently the cost has been reduced tremendously. Resource allocation problems are of

Figure 1. FDM with 9 subcarriers using filters (Intel in Communications, n.d.)

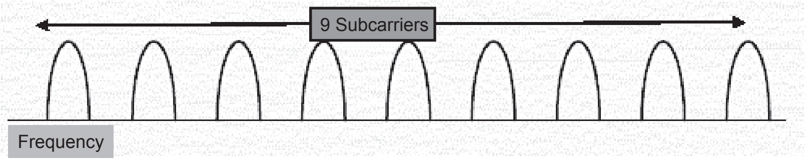
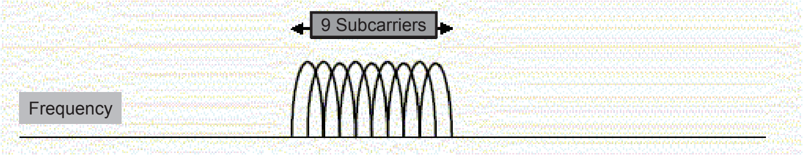


Figure 2. OFDM with 9 subcarriers (Intel in Communications, n.d.)



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

great concern to wireless communication networks dealing with a multiple-user OFDM system (Shen, Andrews, & Evans, 2003).

The major concern is the allocation of subcarriers and power to different users in order to minimize the total transmitted power, increase the total data rate, and maintain an acceptable quality.

BACKGROUND

Mobile and wireless communication networks are growing both in size and complexity and hence the use of OFDM has increased over the past few years. With this tremendous growth, engineers are facing challenges in deciding the allocation of subcarrier, bit, and power to multiple users in an OFDM system. While OFDM-based multi-user systems have been proposed, subcarrier and power allocation for these systems is still under investigation (Kivanc & Liu, 2000). The wireless communication network using OFDM is described as follows (Shen et al, 2003):

- OFDM divides the entire transmission bandwidth into a number of orthogonal subcarriers.
- The set of subcarriers is adaptively assigned to different users. Some subcarriers are inadequate for some users but good for others. Hence by adaptive assignment, we can ensure that all the subcarriers will be used effectively.
- The second issue would be to determine the power level transmitted on each subcarrier in order to minimize the total power transmitted (to reduce cost of transmission) and attain acceptable quality.

Since both the issues are correlated, they must be solved together using a single model (Gross & Karl, 2004).

The problem of optimizing subcarrier and power allocation in a multi-user OFDM system is studied by Kivanc and Liu (2000). Their work explains the importance of the problem and gives a detailed model for the problem. They define two greedy algorithms and claim that they give optimum solutions. The approach is appropriate, but a method such as a greedy algorithm does not guarantee an optimum solution, though the method has less computations and the solution is reached quite fast. The study also includes an application for a small-scale problem using Excel. Using integer programming would yield optimum results as compared to greedy algorithms. A similar approach has been conducted by Wong, Cheng, Letaief, and Murch (1999) and Kivanc and Liu (2000). However, Wong et al. (1999) developed an algorithm for a single-user system, and then they extended the algorithm for a multi-user system. The results are compared to other static allocation schemes and have proved the algorithm to be better than static allocation schemes. Another approach developed by Shen et al. (2003) employs methods like Newton-Raphson and Quasi-Newton. The drawback of the method is the high complexity involved in the computations. Their primary focus is on proving a multi-user OFDM system to be better than the other conventional methods. Their secondary focus is on the optimization technique. Mohanram and Bhashyam (2005) have explored this area and have used suboptimal techniques to solve the problem.

Furthermore, integer programming could be used to solve the problem. In this proposal, the steps required in achieving the optimum solution using integer programming are analyzed. Integer programming gives us the optimal solution, unlike greedy algorithms. However, when the number of design variables increase (a large-scale system model), the time to solve the problem using integer programming is high, and at some point it becomes impossible to solve using integer programming. Thus, an alternate algorithm would be developed for the system, and this algorithm would be tuned to get the same solution as obtained with integer programming. This tuned algorithm can be used for larger models where integer programming could not work. A comparison would be done between the developed algorithm to the existing greedy algorithm in terms of optimality and number of iterations.

When an optimization problem has some variables that can be only integer, then integer programming can be used to achieve the optimum solution (Wolsey, 1998). Evidently, integer programming becomes complex when the number of design variables increase. However, the advantage as compared to heuristic algorithms is that integer programming results in optimum solution. Optimization deals with problems of minimizing or maximizing one or more functions subject to equality or inequality constraints. The major fields of optimization are global, constrained, combinatorial, and multi-objective optimizations (Gen & Cheng, 2000).

- *Global optimization* is maximizing or minimizing a function in the absence of constraints.
- *Constrained optimization* deals with optimizing an objective function subject to equality or inequality constraints.
- *Combinatorial optimization is* used to determine either a permutation or a combination of some items associated with a problem or both. It could also be used to determine the above subject to some constraints.
- *Multi-objective optimization,* as the name suggests, deals with optimizing multiple objective functions simultaneously.

PROPOSED OPTIMIZATION MODEL

A subcarrier and power allocation problem could be classified as a combinatorial optimization problem. The objective of this model is to minimize the total power transmitted and the cost of transmission while maintaining acceptable quality. However, when using integer programming, one of the objectives is made to be a constraint, and hence the problem becomes a constrained optimization problem. In order to develop the algorithm, we must have a complete model for the multi-user OFDM system. The system under consideration is a multi-user OFDM system. Wong et al. (1999) and Kivanc and Liu (2000) have assumed a perfect channel state at both the receiver and the transmitter. The other assumption is that one subcarrier can only be used by one user.

Phase one of this research considers a system with K users and N subcarriers. R_k would be the data rate for the kth user, which means the user k would receive R_k bits. The number of bits of the kth user to the nth subcarrier can be represented by $b_{k,n}$ (Wong et al., 1999). Since our assumption is that one subcarrier can only be used by one user:

for all n, if $b_{k',n} \neq 0$, then $b_{k,n} = 0$ for all $k \neq k'$

$G_{k,n}$ is the magnitude of the channel gain of the *nth* subcarrier as seen by the *kth* user. $P_{k,n}$ is defined as the power spent by the transmitter to transmit to the *kth* user in the *nth* subcarrier, whereas $p(b_{k,n})$ is the signal power required at the receiver side for reliable recovery of $b_{k,n}$ bits (Wong et al., 1999). Therefore:

$p(b_{k,n}) = G_{k,n} \cdot P_{k,n}$ for all k, n

Phase two is to minimize the overall power transmitted. The objective function would be to minimize the overall power transmitted in order to minimize cost of transmission. This could be achieved by minimizing the summation of $P_{k,n}$ for all k, n. The constraints would be that only one user can use one subcarrier, and the total number of bits transmitted to the *kth* user through all subcarriers must equal the data rate. Accordingly, the number of bits for each subcarrier must be assigned. The problem can be seen as an assignment problem where we have to assign the number of bits for each user for each subcarrier minimizing the objective functions. If a particular user is not using a subcarrier, then the number of bits assigned is zero. To summarize the model, we have K users and N subcarriers. The notations used are:

- R_k: Data rate for the kth user, which means the user k would receive R_k bits.
- $b_{k,n}$: The number of bits of the kth user assigned to the nth subcarrier.
- $U(b_{k,n})$: Indicates whether or not subcarrier n is used by user k.
- $U(b_{k,n}) = 0$ only if $b_{k,n} = 0$.
- $G_{k,n}$: The magnitude of the channel gain of the *nth* subcarrier as seen by the *kth* user.
- $P_{k,n}$: The power spent by the transmitter to transmit to the *kth* user in *nth* subcarrier.
- $p(b_{k,n})$: The signal power required at the receiver side for reliable recovery of $b_{k,n}$ bits.

Hence, the objective function is:

$$\min \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{p(b_{k,n})}{G_{k,n}}$$

Subject to:

$$\sum_{n=1}^{N} b_{k,n} = R_k \qquad k=1 \text{ to } K$$

$$\sum_{k=1}^{K} U(b_{k,n}) = 1 \qquad n=1 \text{ to } N$$

$b_{k,n} = 0$ *to M for all k,n*

APPLICATION USING SOLVER

The proposed model is tested using Excel Solver for a different combination of number of users and subcarriers. The number of bits of the kth user assigned to the nth subcarrier $(b_{k,n})$ is marked as the changing variables. Usually it takes a value from zero to the maximum capacity of the channel. Using Solver, it is changed from zero to one and then multiplied by the maximum channel capacity in order to simplify the run. The sum of the number of bits ($b_{k,n}$) for each user is added and verified that the sum equals the data rate for the kth user (R_k).

$U(b_{k,n})$ indicates whether subcarrier n is used by user k or not. It is marked as changing variables and is restricted to be binary. The assumption that one subcarrier can be used by only one user is satisfied by adding $U(b_{k,n})$ for each subcarrier under the condition that the sum remains binary. Hence, as per the proposed model, all three constraints are satisfied. Now in order to achieve the objective function, the proposed method assigns random numbers (between zero and one) to the channel gain. In practice, modeling of the channel is a separate research area. Regardless of the model, the channel gain would always take a value between zero and one. The best channel would have a gain of one, which means that the power transmitted is received in full. On the other hand, the worst channel would have a gain of zero, which means that all the power transmitted is lost.

The signal power required at the receiver side for reliable recovery of $b_{k,n}$ bits ($p(b_{k,n})$) is modeled as square of the number of bits transmitted. Since the relationship between the power and the bits is normally proportional to its square, the proposed method is reasonable. However, the power calculation could be different, but in testing the proposed model, it is irrelevant since the power would always be a real number.

$$p(b_{k,n}) = b_{k,n}^{2}$$

Now, the power transmitted can be calculated by dividing the power required at the receiver side by the channel gain, assuming that the channel gain ($G_{k,n}$) is non-zero. However, it could have a very low value resulting in a huge $P_{k,n}$. The objective function in solver would be to minimize the total power transmitted (sum of all $P_{k,n}$).

$$P_{k,n} = \frac{p(b_{k,n})}{G_{k,n}}$$

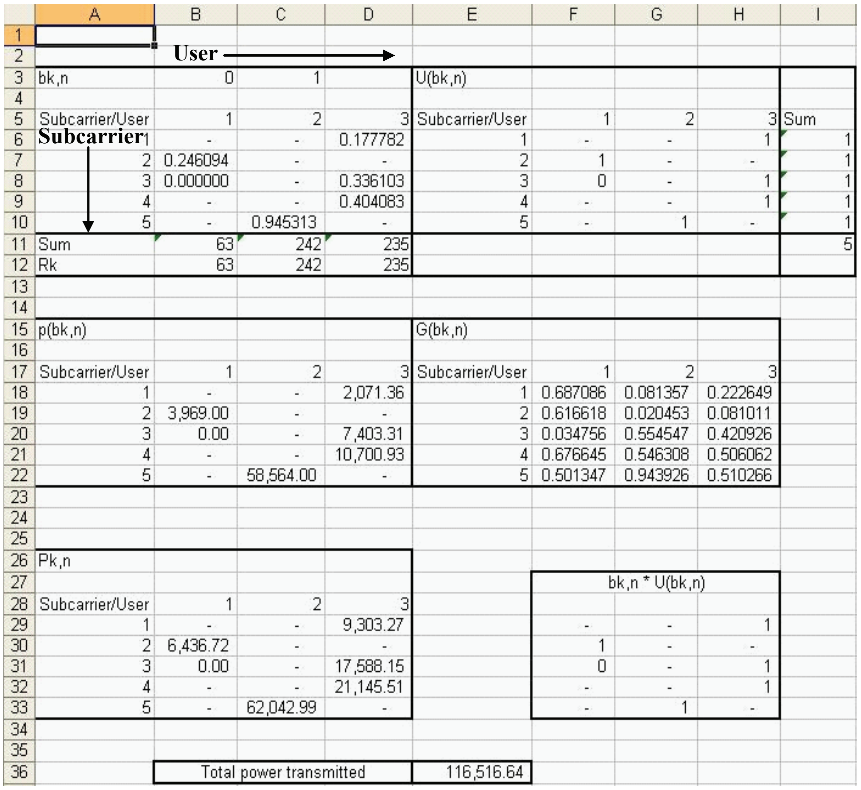
TESTING THE MODEL

A pilot scheme for the model is tested on a small scale using Solver for three users and three subcarriers, and the optimum results were achieved. The three-user model was then run for four and five subcarriers, and optimum results were obtained. When the model was run for four users and four subcarriers, Solver could not find a feasible solution. Since the model returned results for the smaller system, we can conclude that bigger models cannot be tested using Solver because of the restriction of the number of constraints in Solver. Figures 3 and 4 shows the Excel and the Solver screenshot for the 3 × 5 model respectively.

FUTURE TRENDS

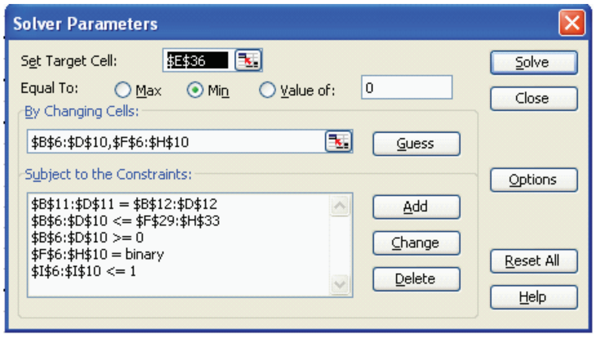
The result obtained using Solver is optimum. However, we cannot use Solver for larger scale models. Hence, an algorithm representing the proposed model using integer programming could be formulated which could be run using programming languages C and MatLab. An alternate approach is that the algorithm could be heuristic. The idea of heuristic algorithms is to "find a good feasible solution quickly" (Wolsey, 1998). Greedy algorithms, genetic algorithms, taboo search, and so forth are some examples of heuristic algorithms. In solving the allocation problem in a multi-user OFDM system, two features could force the use of a heuristic algorithm (Wolsey, 1998):

Figure 3. Excel screenshot for the 3 × 5 model



	A	B	C	D	E	F	G	H	I
1									
2									
3	bk,n	0	1		U(bk,n)				
4									
5	Subcarrier/User	1	2	3	Subcarrier/User	1	2	3	Sum
6	1	-	-	0.177782	1	-	-	1	1
7	2	0.246094	-	-	2	1	-	-	1
8	3	0.000000	-	0.336103	3	0	-	1	1
9	4	-	-	0.404083	4	-	-	1	1
10	5	-	0.945313	-	5	-	1	-	1
11	Sum	63	242	235					5
12	Rk	63	242	235					
13									
14									
15	p(bk,n)				G(bk,n)				
16									
17	Subcarrier/User	1	2	3	Subcarrier/User	1	2	3	
18	1	-	-	2,071.36	1	0.687086	0.081357	0.222649	
19	2	3,969.00	-	-	2	0.616618	0.020453	0.081011	
20	3	0.00	-	7,403.31	3	0.034756	0.554547	0.420926	
21	4	-	-	10,700.93	4	0.676645	0.546308	0.506062	
22	5	-	58,564.00	-	5	0.501347	0.943926	0.510266	
23									
24									
25									
26	Pk,n								
27						bk,n * U(bk,n)			
28	Subcarrier/User	1	2	3					
29	1	-	-	9,303.27		-	-	1	
30	2	6,436.72	-	-		1	-	-	
31	3	0.00	-	17,588.15		0	-	1	
32	4	-	-	21,145.51		-	-	1	
33	5	-	62,042.99	-		-	1	-	
34									
35									
36		Total power transmitted			116,516.64				

Figure 4. Solver screenshot for the 3 × 5 model



Set Target Cell: $E$36

Equal To: ○ Max ● Min ○ Value of: 0

By Changing Cells: $B$6:$D$10,$F$6:$H$10

Subject to the Constraints:

```
$B$11:$D$11 = $B$12:$D$12
$B$6:$D$10 <= $F$29:$H$33
$B$6:$D$10 >= 0
$F$6:$H$10 = binary
$I$6:$I$10 <= 1
```

1. The solution is required quick since the subcarrier and power allocation is done for a dynamic system.
2. The system could become large and complicated where the use of conventional optimization technique is not possible.

Using integer programming to adjust the heuristic method would ensure that the solution is closer to optimal. Genetic algorithm would be a good approach since it has proved to yield faster results than greedy algorithms in some areas. In the proposed optimization problem, the design variable is $b_{k,n}$, whereas we have $p(b_{k,n})$ in the objective function which makes the model nonlinear. However, the signal power required at the receiver side for reliable recovery depends only on the number of bits. For example, to receive one bit, the signal power required at the receiver is the same irrespective of the subcarrier and the user of the incoming signal. Thus, it is possible to remodel the problem, making it linear, and then we can apply linear programming to solve the problem to obtain an optimum solution. The drawback of the linear model is that the system would become more complex. The power required for reliable recovery of bits and the channel gain are not modeled perfectly. Hence, further study in this area would reinforce the proposed model.

CONCLUSION

In this research, a multi-user OFDM system with the challenge of assigning bits and subcarriers to the users while minimizing the overall transmit power is considered. The main purpose is to illustrate the use of integer programming that would simplify the problem in attaining the optimum allocation. Integer programming is used in Solver to achieve optimum solution. However, the limitation of solver is that larger models cannot be tested due to computational restriction in the number of constraints. Other software like C and MatLab could be used to program the model using integer programming. As an alternate solution, the use of genetic algorithms could be considered.

REFERENCES

Gen, M., & Cheng, R. (2000). *Genetic algorithms & engineering optimization.* New York: John Wiley & Sons.

Gross, J., Geerdes, H., Karl, H., & Wolisz, A. (2005). *Performance analysis of dynamic OFDMA systems with inband signaling.* Telecommunication Networks Group, Technical University Berlin, Germany.

Gross, J., & Karl, H. (2004). *Comparison of different fairness approaches in OFDM-FDMA systems.* Telecommunication Networks Group, Technical University Berlin, Germany.

Gross, J., Klaue, J., Karl, H., & Wolisz, A. (2004). Cross-layer optimization of OFDM transmission systems for MPEG-4 video streaming. *Computer Communications, 27,* 1044-1055.

Intel in Communications. (n.d.). *Orthogonal Frequency Division Multiplexing.*

Kivanc, D., & Liu, H. (2000). *Subcarrier allocation and power control for OFDMA.* Department of Electrical Engineering, University of Washington, USA.

Mohanram, C., & Bhashyam, S. (2005). *A sub-optimal joint subcarrier and power allocation algorithm for multiuser OFDM.* Indian Institute of Technology, India.

Shen, Z., Andrews, J., & Evans, B. (2003). *Optimal power allocation in multiuser OFDM systems.* Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, University of Texas at Austin, USA.

Wolsey, L. (1998). *Integer programming.* New York: John Wiley & Sons.

Wong, C., Cheng, R., Letaief, K., & Murch, R. (1999). Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE Journal on Selected Areas in Communications, 17*(10), 1747-1758.

Zheng, K., Huang, L., & Wang, W. (2005). TD-CDM-OFDM: Evolution of TD-SCDMA toward 4G. *IEEE Communications Magazine,* 45-52.

KEY TERMS

4G: Fourth-generation mobile systems.

Frequency Division Multiplexing (FDM): A conventional transmission technique that uses the complete bandwidth.

Heuristic Algorithm: The idea of heuristic algorithms is to "find a good feasible solution quickly," but not necessarily the optimum.

Integer Programming: The method used to solve an optimization problem when all the variables are integer.

Optimization: Deals with problems of minimizing or maximizing one or more functions subject to equality or inequality constraints.

Orthogonal Frequency Division Multiplexing (OFDM): Similar to FDM, but uses less bandwidth.

Resource Allocation: The allocation of the available resources to the user in an efficient way.

Solver: An optimization tool in Microsoft Excel.

Mutual Biometric Authentication

Mostafa El-Said
Grand Valley State University, USA

INTRODUCTION

Cell phone fraud accounts for more than a billion dollars in lost revenue in North America (Rose, 1999; Wingert & Naidu, 2002). Therefore, one of the largest problems in cellular communication systems is the security of cellular phones and the authenticity of cell phone base stations. Caller authentication and voice encryption (CAVE) is the currently used algorithm in cellular systems for authentication and data integrity (Cryptome, 1997; Korzeniowski, 2005; Cellular Technologies, 2006). It was developed by Committee TR45.3 of TIA/EIA under the auspices of the NSA. Gauravaram and Millan (2004) presented two crypto methods to exploit the security vulnerabilities in the CAVE algorithm, and proposed two attack methods that demonstrated that the CAVE is insecure. Another attempt to crack the CAVE is conducted by a research team at the University of California at Berkeley. David Wagner, a member of this research team, announced the CAVE algorithm can be cracked in a matter of minutes or at most hours (Monitoring Times, 1997).

Other attacks on the cellular systems can take place through the RF interface or through the interconnecting wireline networks, including the radio network controller switch and the PSTN network. The RF interface attack is the most severe attack on the cellular systems. Several well-known RF attacks are recognized by the cell phone industry including:

- **Cell Phone Cloning:** The cell phone cloning problem is created because illegal users can capture a cell phone's pair of uniquely assigned numbers—ESN (Electronic Serial Number) and MIN (Mobile Identification Number)—when sending the information to the tower. Then, those illegal users can bill time to a user's account. The cell phone industry solved this problem by encrypting the cell phone number and the pair of ESN and MIN when sending the information to the tower (Hai-Ping Ko, 1996; Landmark Communications, 1996).
- **SMS Messages Flood Attack:** The SMS messages flood attack was a direct result of the shortcoming of 3G cell phone design where the same control channel is used for both call setup and sending SMS messages. An attacker can exploit this vulnerability and use free SMS Web sites to send a large number of anonymous text messages to a cellular phone tower. This could eventually jam up the cellular tower, block any new telephone calls from going through, and result in a denial of service (DoS) attack on the cellular system. The SMS message attack problem was discovered by accident, and some of the European networks have already been jammed when the volume of SMS messaging reached an unexpectedly high level. This problem is still under research (McMillan, 2005).
- **Hash Function Attack:** Gauravaram, McCullagh, and Dawson (2006) investigated several legal and practical implications of attacks against various 128-bit hash functions, and in particular MD5 due to its wide usage. They claim that MD5 can be a single point of failure in various applications. Also, they suggested that new hashing algorithms should be developed in order to avoid new attacks in the future (Gauravaram, Millan, & May, 2004).

This article addresses another severe threat affecting the security of cellular systems, called *pilot aliasing attack*. This problem occurred due to pilot code reuse among different cellular carriers.

In the forward link direction, the base station transmits a dense pack of codes which assists the mobile terminal in performing vital operations such as system synchronization, system acquisition, cell search, and monitoring strong pilots in its serving zone. In general, the CDMA cellular systems have a total of (2^{18} –1) scrambling codes available in the downlink path. These codes are arranged into blocks of 512 coding sets, which are sufficient for a cellular carrier to start deploying a new service. Each cellular tower is allocated one and only one scrambling code, which serves as a unique identifier for the tower and used for cell separation.

These code sets can be reused over and over again within the same carrier's network, and in different carrier networks provided that they will interfere with each other as little as possible (Calhoun, 2003). All 512 cells have the same pilot waveform (same code). They can be differentiated from one another by their pilot signal phase offset. The pilot phase offset is always assigned to the base stations in a multiple of (2^6=64 chips), such as that shown in Figure 1.

Due to the carriers' interoperability nature of the cellular networks, the cell phone may receive two identical pilot signal phase offsets (PN offsets) from two distinct carrier networks at any moment in time. This phenomenon is even worse in the presence of non-homogeneous geography and high-dynamic RF multi-path environments. The mobile

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Base stations pilot code phase offset

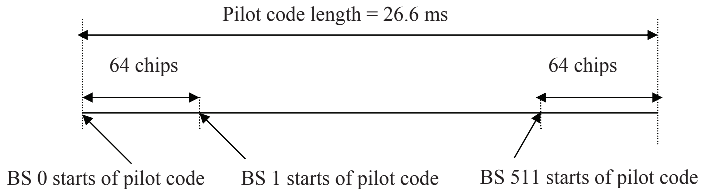
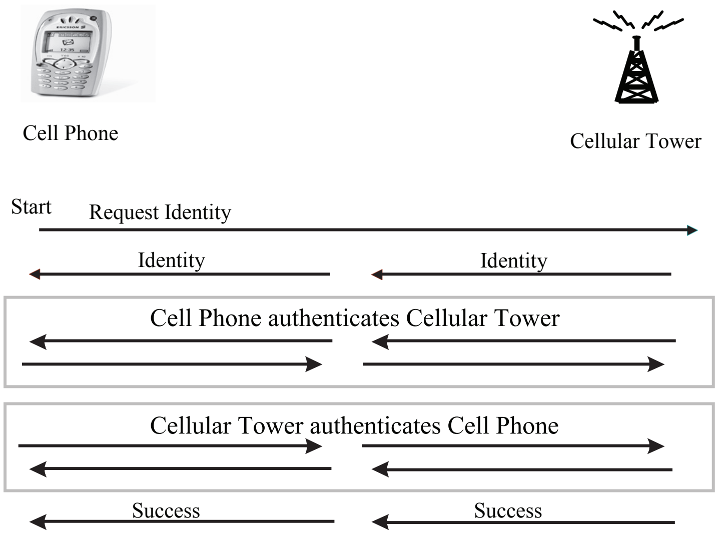

Pilot code length = 26.6 ms
64 chips
64 chips
BS 0 starts of pilot code
BS 1 starts of pilot code
BS 511 starts of pilot code

Figure 2. MBA interaction algorithm


Cell Phone
Cellular Tower
Start
Request Identity
Identity
Identity
Cell Phone authenticates Cellular Tower
Cellular Tower authenticates Cell Phone
Success
Success

receiver became confused and cannot distinguish between the two pilots. One of the pilot signals may be sent from a fake malicious tower that hunts for cell users, which results in *pilot aliasing attack.*

In this article, I use a mutual biometric authentication (MBA-) based solution to allow the cell phone device to authenticate the cellular phone tower before using it as a serving base station. This solution allows the cell phone to detect the existence of pilot aliasing attack. The proposed solution stems from authenticating the cellular tower based on a combined key that consists of the TowerID and the CarrierID's fingerprint. This is a significant step towards creating a robust mutual biometric authentication technique in the cellular system.

PROPOSED SOLUTION

The proposed MBA solution relies on having the tower and the cell phone devices passing each one's credentials before engaging in a communication scenario, such as that described in Figure 2.

The following section describes the proposed MBA interaction algorithm.

MBA Algorithm Assumptions

- The cell phone's Electronic Serial Numb: ESN
- The cell phone's Mobile Identification Number: MIN
- The cell phone's public Number: PPN
- The cellular tower has an ID: TowerID
- The cellular tower Broadcast its PilotID: PilotID
- The CarrierID is kept at the tower and at the cell phone station and acts as a shared secret key between the cell phone and the carrier tower. (See Diagram 1.)

PERFORMANCE MEASUREMENTS AND KEY FINDINGS

To test the applicability of the proposed MBA solution, I use and compare the two hashing algorithms, called MDS and SHA1. To do that, the simulation experiments are carried out 10,000 times. I computed the time that the cell phone

Diagram 1.

Phase I. One-Way Authentication (cellular tower to cell phone station)	
Steps	**Events**
Step 1: At the carrier's tower	The tower hashes its (TowerID+CarrierID+ PilotID) and sends the hash value (Hash1) along with the PilotID and the TowerID in clearText periodically.

Step 2: At the cell phone station	When the cell phone station receives the Hash 1 value, TowerID, and the PilotID, it extracts the TowerID and the PilotID. Then it performs a hash on the (received TowerID + received PilotID + stored CarrierID), called Hash 2.
	The cell phone station compares the two hashes (Hash 1 and Hash 2).
	if (Hash 1 == Hash 2)
	{
	The tower is legitimate.
	} else {
	The tower is a malicious tower and should be avoided }

Phase II. Second-Way Authentication (cell phone station to cellular tower)	
Steps	**Events**
Step 3: At the cell phone station	The cell phone hashes its (ESN+ MIN+ PPN+ CarrierID) and sends the hash value (Hash3) along with the (ESN+ MIN) in clearText to the tower. When the tower receives the Hash 3 value and the (ESN+ MIN), it extracts the (ESN+ MIN) and looks up the PIN value associated with the received (ESN+ MIN). Then it performs a hash on the (received (ESN+ MIN) + retrieved PIN+ CarrierID), called Hash 4.

Step 4: At the carrier's tower	The tower compares the two hashes (Hash 3 and Hash 4).
	if (Hash 3 == Hash 4)
	{
	The cell phone is legitimate.
	} else {
	The cell phone station is bad and should be denied access.
	}

Table 1. Summary of the system performance analysis

	Hash time using (MD-5) in milliseconds	**Hash time using (SHA-1) in milliseconds**
At the Cellular Tower	0.031090	0.035366
At the Cell Phone Station	0.032482	0.036561
Session Hash Time	0.063573	0.071927

station took to hash the incoming message from the tower using MD-5 and SHA-1. The process is repeated at the cellular tower to compute the time that the cellular tower took to lookup the PIN value and hash the incoming message from the cell phone using MD-5 and SHA-1. Table 1 summarizes the simulation results for MD5 and SHA-1.

The results shown in Table 1 comply with the theoretical foundation of the cryptology science, where MD-5 generates a hash value of 128 bits length and the SHA-1 5 generates a hash value of 160 bits length. Therefore, MD-5 should take a shorter amount of time to hash the message compared to SHA-1. The session hash time is very short, especially if we use MD-5 (0.08 ms), and can be reduced if we use high-speed cellular tower stations. The session hash time can be further reduced if we use one-way authentication (cellular tower to cell phone), which guarantees avoidance of the pilot aliasing attack.

FUTURE TRENDS

An effective solution for cell phone security in 3G and 4G wireless systems should investigate the utilization of the asymmetric RSA encryption algorithm as well as the symmetric encryption algorithm. Therefore, the approach proposed in this article can be applied easily to the existing cellular infrastructure.

CONCLUSION

In this article, a mutual biometric authentication has been developed to solve the pilot aliasing attack problem in cellular systems. The proposed solution relies on hashing credentials using MD-5 or SHA-1. The proposed solution is very efficient, especially if we use MD-5, which provides shorter session hash time (0.08 ms) compared to SHA-1 (0.11).

REFERENCES

Agilent Technologies. (2005). Retrieved October 2, 2005, from http://we.home.agilent.com

Calhoun, G. (2003). *Third-generation wireless communications.* Boston: Artech House.

Cellular Technologies. (2006). *How to hack a Motorola GSM phone.* Retrieved April 10, 2006, from http://www.cellular.co.za/technologies/security/how_to_hack_a_motorola_gsm_phone.htm

Cryptome. (1997). *Caller Authentication and Voice Encryption (CAVE).* Retrieved March 5, 2006, from http://cryptome.sabotage.org/cave.htm

Gauravaram, P., McCullagh, A., & Dawson, E. (2006). The legal and practical implications of attacks on 128-bit cryptographic hash functions. *First Monday Journal, 11*(1), 1-2.

Gauravaram, P., & Millan, W. (2004). Cryptanalysis of the Cellular Authentication and Voice Encryption algorithm. *IEICE Electronics Express, 15*, 453-459.

Gauravaram, P., Millan, W., & May, L. (2004). CRUSH: A new cryptographic hash function using iterated halving technique. *Proceedings of the International Workshop on Cryptographic Algorithms and Their Uses* (pp. 28-39).

Hai-Ping, K. (1996). *Attacks on cellular systems.* Retrieved February 22, 2006, from http://seclab.cs.ucdavis.edu/projects/cmad/4-1996/pdfs/Ko.PDF

Korzeniowski, P. (2005). *Cell phone passwords: A weak security link.* Retrieved April 2, 2006, from http://www.technewsworld.com/story/41980.html

Landmark Communications, (1996). *Cell phone scam leader must pay up*. Retrieved April 10, 2006 from http://scholar.lib.vt.edu/VA-news/VA-Pilot/issues/1996/vp960918/09180440.htm

McMillan, R. (2005). *IDG news service. SMS attack could harm cell phones.* Retrieved February 22, 2006, from http://www.pcworld.com/news/article/0,aid,122878,00.asp

Metawave. (2005). Retrieved November 10, 2005. from http://www.metawave.com

Monitoring Times. (1997). *The politics of encryption*. Retrieved March 21, 2006, from http://www.decodesystems.com/mt/97jun/index.html

Rose, G. (1999). Authentication and security in mobile phones. *Proceedings of the Australian Unix User's Group Conference.*

Wingert, C., & Naidu, M. (2002). *CDMA 1xRTT security overview.* White Paper. Retrieved March 4, 2006, from http://www.cdg.org/technology/cdma_technology/white_papers/cdma_1x_security_overview.pdf

KEY TERMS

Caller/Cellular Authentication and Voice Encryption (CAVE): The name of an encryption algorithm used by global cellular phone manufacturers.

Cell Phone Cloning: A mobile phone can be programmed with a stolen or duplicate electronic serial number (ESN) and mobile identification number (MIN). After cloning, both cellular telephones have the same ESN/MIN combination; the cellular systems cannot distinguish between the cloned cellular telephone and the legitimate one.

Cellular Tower's Biometric Authentication: The verification of a tower's identity by means of a physical trait or behavioral characteristic that cannot easily be changed, such as a tower's fingerprint.

Electronic Serial Number (ESN): A unique identification number embedded in a cellular phone by the manufacturer. Whenever a user places a phone call, the ESN is sent to the cellular tower so the wireless carrier's mobile switching office can check the call's validity.

International Mobile Equipment Identifier (IMEI): A unique 15-digit number that acts as a serial number for the mobile handset unit. The IMEI is automatically transmitted by the phone to the network. The cellular carrier can use the IMEI to determine if a mobile unit is stolen or if it is in fault.

Mobile Identification Number (MIN): Uniquely identifies a cellular phone unit within a wireless carrier's network infrastructure. The MIN is assigned by the cellular carrier and often can be dialed from other wireless or wireline networks.

Pilot Aliasing: Term referring to the reception of two identical PN offsets from two distinct CDMA networks, provided that the difference between the pilots' propagation times is less than half of the mobile search window. The mobile receiver becomes confused and cannot distinguish between the two pilots.

New Transaction Management Model

Ziyad Tariq Abdul-Mehdi
Multimedia University, Malaysia

Ali Bin Mamat
Universiti Putra Malaysia, Malaysia

Hamidah Ibrahim
Universiti Putra Malaysia, Malaysia

Mustafa M. Dirs
College University Technology Tun Hussein Onn, Malaysia

INTRODUCTION

As the mobile database permeates into today's computing and communication area, we envision application infrastructures that will increasingly rely on mobile technology. Current mobility applications tend to have a large central server and use mobile platforms only as caching devices. We want to elevate the role of mobile computers to first-class entities in the sense that they allow the mobile user work/update capabilities independent of a central server. In such an environment, several mobile computers may collectively form the entire distributed system of interest. These mobile computers may communicate together in an ad hoc manner by communicating through networks that are formed on demand. Such communication may occur through wired (fixed) or wireless (ad hoc) networks. At any given time, a subset of the computer collection may connect and would require reliable and dependable access to relevant data of interest.

Peer-to-peer (P2P) computing is basically an ad hoc network, and it can be built on the fixed or wireless network. With P2P, computers can communicate directly and share both data and resources. So far, many applications such as ICQ, which allows users to exchange personal messages, and Napster and Freenet, which allow exchange of music files, have taken advantage of P2P technology. However, data management is an outstanding issue and leads directly to the problem of low data availability. Thus, data availability is the central issue in P2P data management. The most important characteristic that affects data availability in the P2P environment is the nature of the network. For the case of the ad hoc network, hosts are connected to the network temporarily. Furthermore, hosts also play the role of router, and they communicate with each other directly without any dedicated hosts. Since there are no dedicated hosts that act as a router, obviously the network connections are prone to get disconnected. Thus, it is difficult to guarantee one-copy serializability since we rely on the mobile hosts, not the fix hosts, in order to communicate with other hosts not reachable directly (Bhargava, 1999). When hosts are disconnected more often and the applications have high transaction rates, the deadlock and reconciliation rate will experience a cubic growth (Bhargava, 1999), the database is in an inconsistent state, and there is no obvious way to repair eventually. For the case of the fixed network, the network connection is relatively stable, but the availability of sufficient computing resource depends on the strategies of replication.

Walborn and Chrysanthis (1996) describe the use of mobile computers in the trucking industry. Each truck has a computer with a satellite or radio link and interacts with the corporate database. Other applications involving remote or disaster areas and military applications have mobile computers forming ad hoc networks without communications with stationary computers. Faiz and Zaslavsky (n.d.) discuss the impact of wireless technologies and mobile hosts on a variety of replication strategies. Distributed replicated file systems such as Ficus and Coda (Agrawal & El Abbadi, 1996) have extensive experience with disconnected operations.

In this article, we consider the distributed database that can make up mobile nodes and the peer-to-peer concept. These nodes are peers and may be replicated both for fault-tolerance, dependability, and to compensate for nodes that are currently disconnected. Thus we have a distributed replicated database where several sites must participate in the synchronization of transactions. The capabilities of the distributed replicated database are extended to allow mobile nodes to plan disconnection, with the capability of updating the database on behalf of the mobile node by using a fixed proxy server to make these updates during the mobile disconnection, once a mobile reconnects automatically, synchronously, and integrates into the database.

By using the notion of planned disconnection MTCO, we present a framework to allow the replicated data of mobile nodes available to access and update at low cost for reading and writing.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

MODEL

A distributed database in a P2P environment consists of a set of data objects stored at different sites (fixed network) and nodes (mobile network) in a computer network. A site (node) may become inaccessible due to site (node) or partitioning failure. No assumptions are made regarding the speed or reliability of the network. Users who interact with the database by invoking transaction must appear atomic: a transaction either commits or aborts (Pitoura & Bhargava, 1995; Dunham and Helal, 1997).

In a replicated database, copies of a data object may be stored at several sites as nodes in the network. Multiple copies of a data object must appear as a single logical data object to the transactions. This is termed as one-copy equivalence and is enforced by the replica control technique. The correctness criteria for replicated database synchronously (fixed network) is one-copy serializability (Dunham & Helal, 1997), which ensures both one-copy serializability and one copy equivalence, while the correctness criteria for replicated database asynchronously (at mobile network) is a timestamp order, which ensures serializability into mobile transactions; a replicated data object may be read by reading a quorum of copies, and it may be written by writing a quorum of copies. The selection of a quorum is restricted by the quorum intersection property to ensure one-copy equivalence: For any two operations o[x] and o'[x] on a data object x, where at least one of them is a write, the quorum must have a non-empty intersection. The quorum for an operation is defined as a set of copies whose number is sufficient to execute that operation.

The environment has two types of networks—the fixed network and the mobile network. For the fixed network, all sites are logically organized in the form of two-dimensional grid structure. For example, if the network consists of 25 sites, it will be logically organized in the form of a 5x5 grid as shown in Figure 1. Each site has a master data file. In the remainder of this article, we assume that replica copies are data files. A site is either operational or failed, and the state (operational or failed) of each site is statistically independent of the others. When a site is operational, the copy at the site is available; otherwise it is unavailable.

The logical structure for fixed and mobile networks is as shown in Figure 1. The circles in the grid represent the sites under the fixed network environment, and *a, b, …,* and *y* represent the master data files located at site s(1,1), s(1,2), …, and s(5,5) respectively. The circles in the oblong shape are sites under the mobile network.

The Technique

In the fixed network, the data file will replicate to diagonal sites according to the DRG technique, discussed in Dircke and Gruenwald (2000), while in the mobile network, the data file will replicate asynchronously at a number of mobile nodes based on the most frequently visited node, and these mobile nodes are not fixed.

For the fixed network, a site s initiates a DRG transaction to update its data object. For all accessible data objects, a DRG transaction attempts to access a DRG quorum. If a DRG transaction gets a DRG write quorum without non-empty intersection, it is accepted for execution and completion, otherwise it is rejected. We assume for the read quorum, if two transactions attempt to read a common data object, read operations do not change the values of the data object. Since read and write quorums must intersect and any two DRG quorums must also intersect, then all transaction executions are one-copy serializable.

For the mobile network, a node will replicate the data asynchronously analogous to the concepts of *checked-out,* proposed in Cetintemel and Kelender (2002). The 'commonly visited node' is defined as the most frequent node that requests the same data at the fixed proxy server (the commonly visited nodes can be given either by a user or selected automatically from a log file/database at each center). This node will replicate the data asynchronously, therefore it will not be considered for the read and write quorums.

Definition 1: Assume that the mobile environment consists of n *mobile nodes. All nodes are labeled* $N_{1,} N_{2,\dots,} N_{n,}$ $1 \leq i \leq n$.

Definition 2: Assume the mobile disconnection consists of a pre-committed transaction and request transaction for maintaining data replicate.

Definition 3: Assume that the limitation amount of data object $X = \Delta$, *where* Δ *calculates from* $\Delta_x = f(x, N_x) = [((1/2) + (0.1*r))\, x/N_x]$.

In Definition 3:

- r = the number of reconnection of replicated data file, and also consider the time by changing the order of data amount: N_x = number of replication and $N_x \leq x$,
- x = value amount of data object x and x > 0, we chose 0.5 to keep some x amount to request transaction,
- pre-committee = transaction process locally at mobile nodes during disconnection depended on the limitation amount of data object x,
- request transaction = transaction process at fixed proxy server because the limit amount of data object x less than the request.

We are using **r** for considering the time by changing the order of data amount, so that in the first disconnection, **r** will start (0), then (0.1***r**); we chose the amount of **(1/2)** in order to save half of the data object amount at fixed proxy server for new disconnection mobile node. After the first

Figure 1. Peer-to-peer environment

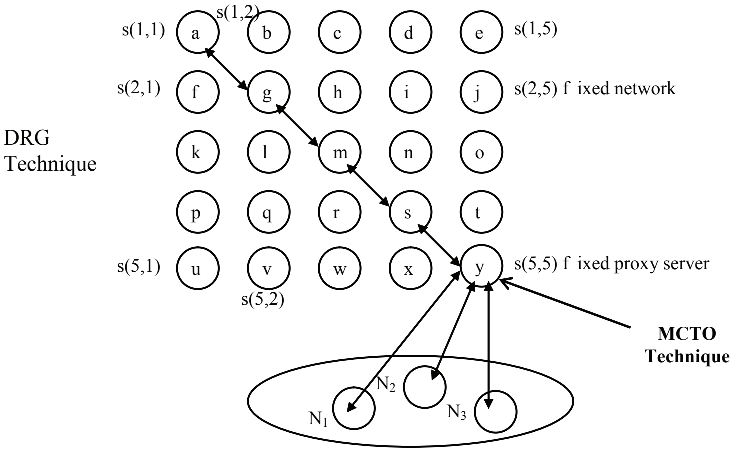
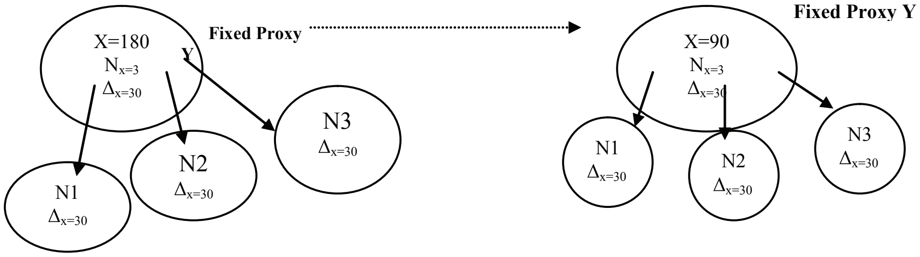


Figure 2. Replicate of data object amount at mobile nodes



disconnection, in any disconnection the fixed proxy server will save less than the half of data object amount depending on the number of (**r***0.1), so that the amount of data object at mobile node will increase depending on the number of **r**, and also the fixed proxy server can allow for the multi-mobile node to disconnect at a different time.

As an example, assume a data object *X* representing the total number of movie tickets, and N_x is the number of replicas of *X* among mobile nodes. Initially $x = 180$ and $N_x = 3$. *X* is replicated at N_1, N_2, N_3. The function that defined in Definition 3 is $\Delta_x = f(x, N_x) = [(0.5 + 0.1r)x/N_x]$; note to keep in first time half of X the amount at fixed proxy server for the request transaction; we are using r to consider the new mobile node wish to disconnect and also from r considering time by changing the orders amount as shown in Figure 2.

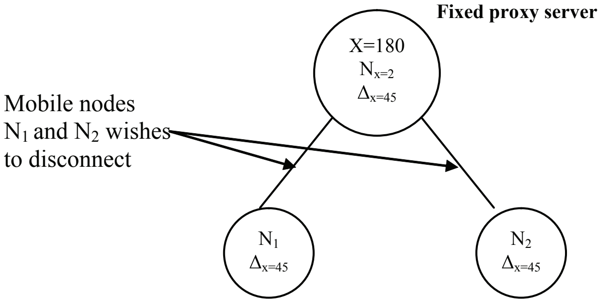
Consider the scenario that N_1 and N_2 wish to disconnect and check-out the data object x, so according to the function in Definition 3 $\Delta_x = 45$, that means mobile nodes N_1 and N_2 will take privilege to maintain locally update less than or equal to the limitation Δ_x and in this case, the mobile node will make pre-commit and when reconnecting to the fixed proxy server will serialize with other transactions depending on timestamp ordering for each transaction and the fixed proxy server will commit as shown in Figure 3.

If the request transaction greater than limitation of Δ_x, in this case, mobile node cannot update data object x locally, when reconnecting it will send the request transaction to the fixed proxy server to process it and send the report to the mobile node, whether abort or commit, depending to the amount of data object X and the transaction arrival time to fixed proxy server.

Consider during mobile nodes N_1 and N_2 disconnect, mobile N_3 wish to disconnect and check-out the same data object X; the fixed proxy allow mobile node N_3 to check-out the data object X according to Definition 3 and will consider the new mobile that wishes to disconnect after N_1 and N_2 as a reconnect mobile, so making r = 1, as shown in Figure 4.

Thus, N_3 will disconnect with privilege update to the data object X less than or equal to the new limitation $\Delta_x = 18$; otherwise, it will send request transaction.

Figure 3. Replication at two mobile nodes



For the multi check-out in the mobile network, if nodes N_1, N_2, N_3 want to disconnect and be able to update the same particular data object, it declares its intention to do so before disconnection and "check-out," or "takes" the object for writing.

This can be accomplished by obtaining a lock on the item before disconnection. An object can be checked out to more than one node at a time. In order to maintain serializability in multi check-out mode, we will use timestamp ordering to serialize the mobile transaction at the fixed proxy server. The fixed proxy server will process two types of transactions—pre-committed transactions and request transactions are executed locally at a mobile node and serialized on the fixed proxy in the order of their timestamp order arrival. The request transaction is a mobile transaction that sends the transaction(s) to a fixed proxy server to update the amount of data because the amount of data at the mobile node is less than the request of transaction and the mobile node cannot update the amount of data locally. So, after the mobile node reconnects to the fixed proxy server, it will send the request transaction to the fixed proxy server asking for an update of the data amount if the amount at fixed proxy server is greater than the request transaction; in this case the fixed proxy server will update the account, commit the transaction, and transfer the committed transaction to the mobile node or abort and send the abort transaction to the mobile node.

The pre-committed transaction is also a kind of mobile transaction, but this transaction can update the data amount at the mobile node; we call this case *pre-committed transaction.* When reconnected, the pre-committed transaction will be transferred to the fixed proxy server to update it, and it will send the committed transaction to the mobile node. But this kind of transaction is different from a request transaction because it can update the data locally.

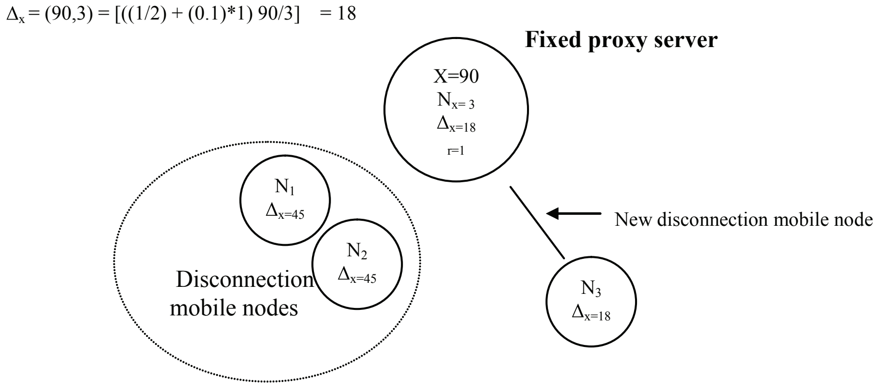
By using timestamp order, the nodes that wish to disconnect, say $N_1, N_2, N_3,$ acquire a write lock on the same object at different times or the same time to update while disconnected. The disconnect procedure is as follows:

1. N_1, N_2, and N_3 tell the nearest "fixed proxy server" from the fixed network to check-out the same data object at any point in time.
2. At the same time, N_1, N_2, and N_3 initiate a transaction to obtain locks on the object depending on the limitation amount that gives equivalence to each mobile node participating.
3. If the transaction is successful, N_1, N_2, and N_3 are disconnects with update privileges on the same object. Otherwise, N_1, N_2, and N_3 try again.
4. N_1, N_2, and N_3 can make local transaction and pre-commit during the limitation amount given to them; if the transaction is greater than limitation, the mobile node will send a request transaction to the fixed proxy server when connected.

The reconnect procedure is as follows:

1. When N_1, N_2, and N_3 reconnect, it contacts the fixed proxy server from the fixed network.

Figure 4. New mobile node wishing to disconnect



2. N_1, N_2, and N_3 will transfer the pre-commit transaction and request transaction to the fixed proxy server. The pre-committed transaction will transfer the transaction(s) that were updated during mobile node disconnection, while the request will transfer the transaction(s) that were not updated during mobile disconnection because the transaction request was greater than the data amount at the mobile node.
3. The fixed proxy server applies the DRG to replicate to diagonal sites at the fixed network only.
4. N_1, N_2, and N_3 release the corresponding lock.
5. The fixed proxy server sends the data with the latest updated version to N_1, N_2, and N_3.

In order to preserve correctness, it must be possible to serialize all of the transactions executed by N_1, N_2, and N_3 during disconnection at the point in time of reconnection. This can be done if:

1. Objects not write locked by transaction at disconnect time are treated as read-only by N_1, N_2, and N_3 during disconnect.
2. The object process at the fixed proxy server depends on the transaction time arrived.
3. Timestamp ordering serializes all mobile transactions.

This will guarantee serializability because each transaction at a disconnected node respects the timestamp ordering.

Since the data file is replicated to only the diagonal sites at the fixed network, then it minimizes the number of database update operations, misrouted and dropped out calls. Also, sites are autonomous for processing different query or update operations, which consequently reduces the query response time. The data files are proved in Dircke and Gruenwald (2000).

PROOF OF CORRECTNESS AND PROPERTIES

In this section, we prove the correctness of our new technique for requested transactions; multi check-out timestamps order—that is, the MCTO technique—ensures serializability for request transaction. In the correctness proof, we prove that all committed schedules *S* produced by the requested transaction MCTO technique are serializable. In other words, the serialization graph *SG (S)* does not contain any cycle (Forman & Zahorjan, 1994). The SG is a directed graph $G = (N,E)$ that consists of a set of nodes $N = \{T_1,T_2,\ldots,T_n\}$ and a set of directed edges $E = \{e_1,e_2,\ldots,e_n\}$. There is one node in the graph for each transaction T_i in the schedule. Each edge e_i in the graph is of the form $(T_i \rightarrow T_j)$, $1 \leq i \leq n$, $1 \leq j \leq n$, where T_i is the initial node of e_i and T_j is the end node of e_i. Such an edge is created if one of the operations in T_i appears in the schedule before some conflicting operation in T_j.

Serializability is a widely used correctness criterion of database consistency (Unland & Schlageter, 1992). It is more difficult to maintain serializability of mobile transactions than that of traditional distributed database transactions.

Lemma 1

Let T_i and T_j be two committed transactions in a schedule S produced by the MCTO technique. If there is an edge $T_i \rightarrow T_j$ in SG (S), then $TS(T_i) < TS(T_j)$.

Proof

If there is an edge $T_i \rightarrow T_j$ in SG (S), there must exist one or more conflicting operations of one of the following types on data object x.

In the MCTO technique, the serialization order among transactions may not be the same as the chronological order of transaction commitment T_j may commit before T_i does. Therefore, two possible cases of commit order must be considered.

Case 1: $r_i[x] \rightarrow w_j[x]$

T_i commits (read) before T_j reaches fixed proxy server. When T_j reaches the fixed proxy server, for $w_j[x]$, the timestamp interval of T_j will be adjusted: $TI(T_j) := TI(T_j) \cap [RTS(x), \infty]$. It ensures that the final timestamp interval of T_j, $TS(T_j)$, which is obtained by adding a sufficiently small value to the lower bound of $TI(T_j)$, is greater than RTS (x), which is equal to or greater than $TS(T_i)$ or T_j aborts if the interval shuts out. Therefore, if T_j can be committed, then $TS(T_i) \leq RTS(x) < TS(T_j)$, thus $TS(T_i) < TS(T_j)$.

Case 2: $w_i[x] \rightarrow r_j[x]$

T_i commits(write) before T_j reaches fixed proxy server. This case happens when T_i commits before T_j reads x. When T_j reads x, the timestamp interval of T_j will be adjusted: $TI(T_j) \cap [WTS(x), \infty]$ where WTS (x) is equal to $TS(T_i)$. It ensures that $TS(T_j)$ is greater than $TS(T_i)$ or T_j aborts if the interval shuts out. Therefore, if T_j can be committed, then $TS(T_i) < TS(T_j)$.

Theorem 2

If S is a committed schedule produced by the MCTO technique, then S is serializable.

Proof

Consider there is an edge in SG (S), $T_i \rightarrow T_j$, then there must exist conflicting operations $p_i[x]$ and $q_j[x]$ in S, such that $p_i[x]$ precedes $q_j[x]$. Hence, by Lemma 1, TS $(T_i) <$ TS (T_j). If a cycle $T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n \rightarrow T_1$ existed in SG (S), then by induction, TS $(T_1) <$ TS (T_1). This is a contradiction. Therefore, SG (S) is not cyclic and thus S is serializable.

Lemma 3

All transactions that are committed by the DRG technique are also committed by the MCTO technique.

Proof

Assume a transaction T_i is committed by the DRG technique but is aborted by the MCTO technique. Let RC (T_i) be the set of recently committed transactions that are committed between the time when T_i starts its execution and the time at which it reaches the fixed proxy server phase. Since T_i is committed by the DRG technique, it must be true that RS $(T_i) \cap$WS $(T_{RC}) = Ø$ for each transaction $T_{RC} \equiv$ RC (T_i). If T_i is aborted by the MCTO technique, it implies that TI (T_i) shuts out and T_i must read some data objects written by at least some transaction $T_{RC} \equiv$ RC (T_i). Hence, RS $(T_i) \cap$ WS $(T_{RC}) \neq Ø$, which is in contradiction to the assumption, RS $(T_i) \cap$ WS $(T_{RC}) = Ø$, that T_i is committed by the DRG technique. Thus, Lemma 3 follows.

Lemma 4

All transactions that are aborted by the MCTO technique are also aborted by the DRG technique.

Proof

Assume a transaction, T_i is aborted by the MCTO technique but is committed by the DRG technique. Since T_i is aborted by the MCTO technique, it must be true that TI (T_i) shuts out and T_i read some data objects written by at least some transactions $T_{RC} \equiv$ RC (T_i). That is, RS $(T_i) \cap$ WS $(T_{RC}) = Ø$ where $T_{RC} \equiv$ RC (T_i). If T_i is committed by the DRG technique, it must be true that RS $(T_i) \cap$ WS $(T_{RC}) = Ø$ for each transaction $T_{RC} \equiv$ RC (T_i), which is in contradiction to the assumption that T_i is aborted by the MCTO technique. Thus, Lemma 4 follows.

Theorem 5

The set of transactions committed by the DRG technique is the subset of transactions committed by the MCTO technique.

Proof

The theorem follows directly from Lemma 3, Lemma 4, and Example 1 in which T_2 is aborted by the DRG technique, but is committed by the MCTO technique.

CONCLUSION

Transaction management in mobile databases has the same behaviors as those in multi-database systems in many aspects. Many approaches in multi-database systems can be extended to a mobile multi-database environment. The differences in mobile databases are that transactions in mobile databases have mobility and long-lived nature. In this article we have developed a mobile transaction model that captures data and movement nature of mobile transactions. This model is based on multi check-out. The model describes a mobile transaction *Management by Timestamp Order*. Based on this mobile transaction model, we have presented a serializability theory of mobile transactions in mobile databases. In addition, the correctness of the technique was discussed.

REFERENCES

Agrawal, D., & El Abbadi, A. (1996). Using reconfiguration for efficient management of replicated data. *IEEE Transactions on Knowledge and Data Engineering, 8*(5), 786-801.

Agrawal, D., & El Abbadi, A. (1990). The tree quorum protocol: An efficient approach for managing replicated data. *Proceedings of the 16th International Conference on Very Large Data Bases* (pp. 243-254).

Alonso, R., & Korth, H. F. (1993). Database system issues in nomadic computing. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 388-392).

Bell, D. A. (1992). *Distributed database systems.* Boston: Addison-Wesley.

Bernstein, P.A., Hadzilacos, V., & Goodman, N. (1987). *Concurrency control and recovery in database systems.* Boston: Addison-Wesley.

Bhargava, B. (1999). Concurrency control in database system. *IEEE Transaction Knowledge and Data Engineering, 11*(1), 3-16.

Bright, M. W., Hurson, A. R., & Pakzad, S. H. (1992). A taxonomy and current issues in multidatabase systems. *IEEE Computer, 25*(3), 50-60.

Buretta, M. (1997). *Data replication: Tools and techniques for managing distributed information.* New York: John Wiley & Sons.

Dircke, R., & Gruenwald, L. (2000, December). A pre-serialization transaction management technique for mobile multi-database. *ACM Mobile Networks and Applications, 5*, 311-321.

Dircke, R. A., & Gruenwal, L. (1998). Nomadic transaction management. *IEEE Potentials, 17*(2), 31-33.

Dunham, M. H., & Helal, A. (1997). A mobile transaction model that captures both the data and the movement behavior. *ACM/Baltzer Journal on Special Topics in Mobile Networks and Applications, 2,* 149-162.

Elmasri, R., & Navathe, S. B. (2000). *Fundamentals of database system* (3rd ed.). Boston: Addison-Wesley.

Faiz, M., & Zaslavsky, A. (n.d.). *Database replica management strategies in multidatabase systems with mobile hosts.*

Forman, G. H., & Zahorjan, J. (1994). The challenges of mobile computing. *IEEE Computer, 27*(4), 38-47.

Garcia-Molina, H. (1988). Node autonomy in distributed systems. *Proceedings of the International Symposium on Databases in Parallel and Distributed Systems* (pp. 158-166).

Garcia-Molina, H., & Barbara, D. (1985). How to assign votes in a distributed system. *Journal of the ACM, 32*(4), 841-860.

Gary, J., & Reuter, A. (1993). *Transaction processing: Concepts and technique.* San Francisco: Morgan Kaufman.

Gruenwald, L., & Banik, S.M. (2001, September). A power-aware technique to man age real-time database transactions in mobile ad-hoc networks. *Proceedings of the 4th International Workshop on Mobility in Database and Distributed Systems, part of the International Conference on Database and Expert Systems Applications (DEXA)*.

Holliday, J., Agrawal, D., & El Abbadi, A. (2003). Epidemic algorithms for replicated databases. *IEEE Transactions on Knowledge and Data Engineering, 15*(5), 1218-1238.

Holliday, J., Agrawal, D., & El Abbadi, A. (2000, February). Exploiting planned disconnection in mobile environments. *Proceedings of the 10th IEEE Workshop on Research Issues in Data Engineering (RIDE2000)* (pp. 25-29).

Holliday, J., Agrawal, D., & El Abbadi, A. Disconnection modes for mobile databases. *Journal of Wireless Network, 8,* 391-402.

Kung, H. T., & Robinson, J. T. (1981). On optimistic methods for concurrency control. *ACM TODS, 6*(2).

Li, C. (2000, May). *Replication protocol for mobile data access system—an approach under summary schemas model.* Master's thesis, Department of Computer Science and Engineering, The Pennsylvania State University, USA.

Lim, J. B., Hurson, A. R., & Kavi, K. M. (1999). Concurrent data access in a mobile heterogeneous system. *Proceedings of the 32nd Annual Hawaii International Conference on System Sciences.*

Mat Deris, M. (2001). *Efficient access of replication data in distributed database systems* (p. 38). PhD thesis, University Putra, Malaysia.

Mustafa, M. D., Nathrah, B., Suzuri, M. H., & Abu Osman, M. T. (2004). Improving data availability using hybrid replication technique in peer-to-peer environments. *Journal of Interconnection Networks, 5*(3), 299-312.

Ozsu, M. T., & Valduriez, P. (1999). *Principles of distributed system* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Pitoura, E., & Bhargava, B., (1995). Maintaining consistency of data in mobile distributed environments. *Proceedings of the IEEE Workshop on Mobile Systems and Applications.*

Ramasubramanian, R. (1998, August). *A survey of replication issues and strategies in mobile and multidatabase environments.* MEng technical paper, Department of Computer Science and Engineering, The Pennsylvania State University, USA.

Skeen, D. (1985). Determining the last process to fail. *ACM Transactions on Computer Systems, 3*(1).

Unland, R., & Schlageter, G. (1992). A transaction manager development facility for non-standard database systems. In A. K. Elmagarmid (Ed.), *Database transaction models for advanced applications* (pp. 400-466). San Francisco: Morgan Kaufmann.

Walborn, G., & Chrysanthis, P. (1996). Transaction processing in promotion. *Proceedings of the ACM Symposium on Applied Computing.*

Wolfson, O., Jajodia, S., & Huang, Y. (1997). An adaptive data replication algorithm. *ACM Transactions on Database Systems, 22*(2), 225-314.

Next-Generation Mobile Technologies

Chor Min Tan
British Telecommunications (Asian Research Center), Malaysia

Choong Ming Chin
British Telecommunications (Asian Research Center), Malaysia

Moh Lim Sim
Multimedia University, Malaysia

INTRODUCTION

Mobile communications and the Internet have experienced rapid and largely unexpected growth during the last decade of the 20th century. Consequently, the mobile triple-play services (voice, video, data) will be the major demand drivers for the emerging 21st century networks. In this view, the convergence of mobile communications, multimedia, and the Internet would produce innovations, novel applications, and new services that would not otherwise be possible (ITU, 2002). Rapid technological innovations have successfully characterised and encouraged the evolutions of various mobile access technologies, leading to the development of various interworking solutions across heterogeneous networks, as well as the transformation of the market structure and business model. The interworking of heterogeneous networks and technologies is where convergence is really exploding, and it is part of the Internet and multimedia movement that has shaped the future world more than the invention of the automobile (Vanjoki, 2005).

This article aims to present the trends of convergence of different mobile technologies in order to meet the requirements in the provisioning of wireless broadband services, followed by an assessment of its impact to the telecommunications community. The discussions will include an insight into the potential collaborations between various technologies, as well as an exploration of possible technical and economical challenges along the convergence trail.

BACKGROUND

Towards the 21st century, there is a strong need to bring the desktop experience to a mobile environment that allows freedom of movement at any speed while connected to the best network. While the advances of Internet have motivated the research and development (R&D) of wireless technologies that adopt network architecture based on Internet protocol (IP), migration of telephony from circuit-based to voice-over IP (VoIP) might transform the traditional telephony business and change the economics of carrying voice traffic. The emergence of various innovative technologies that support VoIP applications would allow service providers and operators to penetrate into the voice market that was previously inaccessible. In addition, the increased demands for mobile entertainment have motivated the evolution of revenue-producing and bandwidth-hungry multimedia applications such as the digital TV, mobile TV, and Internet (or IP) TV [e.g., TV data, TV telephony (Nortel, 2005)]. Mobile entertainments enable an increasingly compelling content offering as well as new methods to deliver video programming and advertisements to mobile consumers, introducing new business models for both live and on-demand video content. These enhancements are changing how telecommunication networks can be used to enrich subscribers' quality of living as well as to generate additional revenue streams for content providers and network operators. It can be foreseen that the emerging technologies would support various high-end communications, where the requirement of full support for mobile triple-play services is of the ultimate importance.

Following that, various new broadband technologies have been (or are being) developed to address the demands of high-definition multimedia services in future wireless environments. These technologies include Wi-Fi[1], WiMAX[2], 3G evolutions (HSDPA[3]/HSUPA[4], and EV-DO[5]), FLASH-OFDM[6], Mobile-Fi, DVB-H[7], DMB[8], FLO[9], ISDB-T[10], and so forth, to name a few most popular ones. Wireless operators have been spending billions to upgrade their infrastructure and networks, especially during the period of analog-to-digital transition. Today, in order to counter the threats from new technologies like WiMAX particularly in terms of coverage range and data rates, Wi-Fi players are now deploying the mesh topology to enlarge their hotspot coverage range for the deployment of citywide wireless network (MuniWireless, 2006). The Enhanced Wireless Consortium (EWC, 2006) has also published the high-throughput (over 100 Mbps) specifications for the physical and medium access control layers of the Wi-Fi system. It is likely that the IEEE 802.11n standard will be based on the EWC proposals. Further, the IEEE 802.11e standard has also been developed to improve

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the quality of service (QoS) of the legacy Wi-Fi system in provisioning multimedia services, which in turn is able to challenge the QoS-guaranteed technologies like WiMAX and evolved 3G from securing a leading market position in multimedia last mile delivery.

The high demand for mobile broadband and the rapid adoption of multimedia applications have motivated service providers and operators to deploy various networks to cater to the needs of tomorrow. To some extent, some may believe that these technologies are competing with each other, as many of them have been developed or customised to support a same application category or service offering. Technological wars are likely to be waged over the next few years, especially between WiMAX and evolved 3G, as well as between DVB-H and FLO, although no technology is yet in a leadership position. Moreover, as mobile operators have made vast investments in 3G spectrum, they do not wish to see other technologies penetrating into their cellular territories, cannibalising their revenue and profits, especially in the voice markets. However, some parties actually believe that these technologies could complement and strengthen each other in many aspects (Tan et al., 2006). These technologies can run alongside one another bridging the gaps of applications and services rather than overlapping with neighbouring technologies. With careful planning, different technologies can be deployed to demonstrate the interworking and convergence of various service functionalities, and hence interacting in ever more important roles.

The common goals and functionalities of these technologies include the supports for mobility, QoS, ubiquitous access (large coverage range), low error rates, high-speed connections, seamless handover between cells or base stations, and high capacity for simultaneous users. Although these technologies have different capabilities and marketplaces, the convergence of different networks is not far off as most advanced technologies are (being) developed to address data-centric and IP-based applications. The convergence phenomenon is not just in terms of similarity in functionalities, it is also about several networks coming together to further enrich user experience for "always-on" data and multimedia content delivery. This must also be complemented by enhanced user device capabilities, such that a basic device like a mobile phone can become an all-in-one telecommunications, media and computer handheld machine. Carrier operators and service providers all recognise that to sustain their business long-term, they need to devise the right mix of technologies at the right price to lock in subscriber loyalty, entice new customers and increase their average revenue per unit. The shift to new generation networks for mobile entertainments will take this phenomenon even further, and will essentially involve convergence and interoperability in terms of harmonisation and consolidation of technological strengths, functionalities, service offerings, and mobile broadband market segmentation.

CONVERGENCE PHENOMENA

The activities along the convergence trails have revolutionised and created many new technologies, and the wireless R&D community has already begun to glue the broadband Internet, high-definition multimedia, and mobile communications onto a more common ground. On a technical level, the viability of next-generation networks will rely on continued efforts towards the provisioning of ubiquitous access, definite guarantees on QoS supports, high transmission speeds for downlink and uplink, and evolution to IP-based core networks. Despite the availability of various wireless technologies, it can be observed that most technologies have been developed (or upgraded) to fulfill the above requirements, that is, the convergence of technical trends. In order to offer ubiquitous access, Wi-Fi coverage areas are being expanded to form a metropolitan network via the mesh topology. The process is being accelerated with the aid of WiMAX that supports point-to-multipoint architecture and serves as wireless backhaul for several Wi-Fi access points (Intel, 2004a). The rapid growth of VoIP and multimedia applications has placed QoS issues to be part of the network selection criteria, and hence all new technologies in the future would somehow incorporate QoS supports into the system.

In terms of radio access, the design of the latest wireless air interface can be seen to be gradually converging to the use of orthogonal frequency division multiplexing (OFDM) (Intel, 2004b) technology as the preferred radio transceiver technique. The phenomenon can be observed in various systems such as Wi-Fi, WiMAX, FLASH-OFDM, DVB-H, FLO, ISDB-T, and so forth. Further, it is anticipated that future releases of 3G standards would adopt OFDM technology, where the move can be seen in various proposals amongst the industries. These include the high-speed OFDM packet access (HSOPA) by Nortel (Duplessis, 2005) and the Super 3G vision [or long-term evolution (LTE), 3.99G] by NTT DoCoMo (3G Mobile, 2005). In addition, in order to address the problems of bandwidth and transmission speed limitations, most new wireless technologies have been adopting multiple-input multiple-output (MIMO) and smart antenna techniques. It can be foreseen that future consumer devices would employ multiple antennas as one of the key solutions to boost data rates, transmission reliability, and spectral efficiency.

At the service level, convergence between various wireless networks is already happening through new technologies such as unlicensed mobile access (UMA), interworking and interoperability solutions for co-existing networks ("always-on" communications), mobile TV, interactive end-user applications via alternative networks, and so forth. Currently, this form of convergence is at the height of its technology and is strengthening with further developments and innovations. With the advent of VoIP applications, the threats of mobile operators suffering from falling revenues for voice services have prompted them to strive to encourage the use of convergent technology such as the UMA. Cellular community has already started to collaborate with the wireless local area network (WLAN) operators (public and private) to enable subscribers to make voice calls using UMA-enabled mobile phones (e.g., BT Fusion, DT Dual-Phone, KT One-Phone) within the WLAN (mainly Wi-Fi-based) environments, while at the same time providing transparent and single-billing facility. The increased demands for low-cost voice applications would assemble several networks of different technologies towards the convergence into a common service platform.

On another hand, it is envisioned that most wireless users will hook up to applications offering triple-play services of multimedia, broadband Internet and telephony by the end of the decade. The proliferation of mobile entertainments is also following a similar trajectory as that of the VoIP, where the mobile phones will be used to receive live TV programmes. Exploiting high-definition multimedia applications across the cellular core will require high levels of investment, possibly higher than the migration from 2G to 3G and beyond. This is mostly due to the inherent limitations of 3G networks for carrying "unicast" video traffics (highly bandwidth-intensive). In order to avoid overloading 3G traffics with videos, technologies like DVB-H, FLO, and DMB have been developed to allow mobile networks to broadcast live TV shows to consumer devices (e.g., mobile phones) nationwide. This capability of convergence has incorporated both on-demand contents from 3G networks and live programming over the digital broadcast channels. It enables the handset to transcend from traditional real-time voice communication gadget to an instrument facilitating a greater interaction between telecommunications, multimedia, and information technology, revolutionising the handset to be a new generation pocket TV. The lesson learned from the success or failures of new technologies is that timely availability of consumer devices with simple and convenient interfaces will be the key to future service development, revenue generation, and establishment of leading market position. Devices capable of supporting high-quality multimedia combined with additional functionalities, such as VoIP, broadband Internet, interactive gaming, and location technologies, will further enhance the user experience.

Moreover, the realisation for "always-on" radio communications is around the corner. This can be seen in the Network Working Group of the WiMAX Forum that is currently investigating the WiMAX-3GPP interworking solutions, as well as the IEEE 802.21 Task Group (IEEE 802.21, 2006) that is looking into seamless handover solutions across heterogeneous networks. This convergence scenario would eventually encompass complimentary and alternative network technologies, such as UMA and fixed-mobile convergence, where advanced mobility and radio resource management would be considered in their global context. The collaborations between several technologies allow mobile users to stay connected with the best network while roaming from one base station to another. For example, the video telephony applications can be delivered via 3G networks, while heavy files uploading or downloading can be accomplished simultaneously via global broadband access networks like WiMAX and Wi-Fi.

Furthering this, the collaborations between heterogeneous networks have also allowed different systems to coexist and colocate at a same site. This phenomenon is now actively promoted within the Open Base Station Architecture Initiative (OBSAI, 2006), especially regarding WiMAX and 3G cellular systems. From the perspective of OBSAI's open specifications, vendors can now build multimode base station components and module cards supporting the standard hardware of several variants of cellular technologies and WiMAX. This change, impelled by market's demands, has prompted different networks to band together to prevail. It has shaped the way telecommunication services are provided, enabling the coexistence and interoperability across heterogeneous networks, and hence gradually driving several technologies towards a converged marketplace. The future will see the service providers or operators offering more functions comprising more technologies and applications. It should be noted that despite the on-going competition amongst different wireless technologies, none of them should be considered as a substitute for others, as most technologies complement one another rather than competing with each other (Tan et al., 2006).

CHALLENGES

While there are numerous benefits of having a converged communications network, most of the convergence phenomena discussed above represent only a theoretical observation and finding under ideal circumstances. Along the convergence trail, several technical, regulatory, and business issues are still unresolved. Due to the existence of a plethora of wireless technologies, each having its own relative strengths and weaknesses, the crowded technical environment might result in a highly fragmented, unpredictable, confused, and more competitive marketplace. This is especially so as most operators, content and service providers would like to recoup their investments on current technologies (3G in particular) before migrating to yet another new network architecture aiming for convergence. Before 3G communications were deployed few years ago, some parties have been talking about "4G," which is still in principle a buzz word as nothing solid has been specified so far. Initially, a 4G system is supposed to partly address the convergence issues, as some parties have been looking into software defined radio (SDR, 2006) as the core for next-generation systems. It is a pretty ambitious type of network, as ultimately it intends to integrate several networks together using software control so that users can move between cells of different types of radio technologies seamlessly. While the existing technologies (e.g., HSDPA, EV-DO, WiMAX, Wi-Fi mesh) are still being deployed, not much effort has been placed into the so-called "4G," as this requires huge amounts of innovations and investments. It seems that the entire industry is not fully prepared for the convergence yet, not until at least they have firmly secured individual marketplace for the current technologies.

Another greatest challenge will lie with the power consumption issue. It is envisaged that the converged communication platform in future would require multiple processing and additional communication elements, further draining the battery life of the mobile devices. Most mobile devices available today fail to offer sufficient battery life for continuous intensive usage of more than 3 hours. As digital broadcast TV moves to mass market in near future, it is possible that the battery life of a handset to be measured in TV-time, not just talk-time and standby-time (Texas Instruments, 2005). Advanced power management solutions must therefore be in place. Moreover, the mobile device manufacturers and silicon solution providers must be able to balance the increased functionalities with what consumers have been expecting—small designs, reliable communication services, and ease of use. These include the considerations of bigger and higher contrast display screen for the handset, as well as stylish antenna design. Compact size antenna must be able to receive live TV broadcast and wireless communication services (e.g., 3G, WiMAX) in all on-the-go scenarios, whether travelling on a train through a tunnel or sitting still at a desk in an interior office of a high rise (Texas Instruments, 2005). Handset manufacturers will have to satisfy all these requirements while lowering the production costs. The availability of appropriate mobile devices at affordable prices will be a prerequisite for users.

In addition, multimode technology in a convergence scenario means the ability to perform seamless handover between different types of radio access technologies. Towards this end, there are significant software, billing, carrier and enterprise interoperability challenges. The next-generation mobile communications will not be driven by a single entity or organisation. It requires a tremendous number of partnerships and a robust ecosystem in order to fully exploit the ultimate capabilities of all wireless systems. Given the sweeping changes in the world of technology, it is really going to require multiple standardisation bodies, corporations, industries, and government entities to come together to drive open standard-based interoperability, interworking, and opportunity to enable triple-play merged services. Simple and transparent billing models across several parties are required, taking into account the differences between voice and data services and the growing importance of contents.

Further, the transition from circuit-switched networks to all-IP networks faces different challenges of packet acceleration, traffic management, data integrity, security, and QoS. Next-generation security will have multiple elements, much more than just delivering encrypted traffic at faster rates across the converged networks. It is also about denial of service attacks and digital right management. The improved QoS support is required to identify and prioritise data packets in order to offer uninterrupted service flows while accessing multiple networks simultaneously (Perkins, 2005). For example, the mobile device must be able to receive uninterrupted live TV shows in real-time, while simultaneously receiving voice calls from multiple parties and downloading large data files from office servers.

Clearly, the mobile industry will have to look into these challenges, as solving these potential problems is at the forefront of convergent mobile technologies. With that, it may be difficult to expect the commercial fruit of the converged mobile communications within the next 10 years or so.

CONCLUSION

This article has discussed the trends, future scenarios, and challenges regarding the convergence of next-generation mobile communications, where the collaboration and interworking between heterogeneous networks and technologies have been specially highlighted. It is shown that the convergence phenomenon is not just in terms of similarity in functionalities or technical requirements, it is also about several networks of different technologies and standards merging together to further enrich user experience in a more efficient manner. It is expected that the integration of mobile entertainments, Internet, and wireless technologies would give birth to a whole new family of services and applications, as well as opportunities for revenue generation. Towards the convergence path, a new era of pervasive computing is dawning with huge implications for consumer's lifestyles and values, where each may own some sort of miniaturised mobile communication devices (e.g., multimode pocket computer with multimedia and telecommunication functionalities) that are always best connected. This emerging trend is mostly driven by wide range of requirements, where the appetite for seamless mobility with an all-in-one mobile device is of the ultimate importance.

However, the convergent marketplace is still in its infancy, and a realistic business case that is both commercially available and overwhelmingly profitable is still yet to be defined. It is anticipated that business models based on partnerships will be the cornerstones for the continued success of the convergent mobile industry to generate opportunities for all players and foster innovations. A significant level of investment, support, understanding, and cooperation between various parties is needed in order to satisfy various requirements of a converged next-generation mobile network. Proper measures to resolve the foreseen challenges are also vital to achieving the ultimate success of the converged next-generation wireless architecture.

REFERENCES

3G Mobile. (2005, January 19). Super 3G reveals long-term failure of WCDMA. *3G Mobile*.

Duplessis, P. (2005, July). HSOPA: Exploiting OFDM and MIMO to take UMTS beyond HSDPA/HSUPA. *Nortel Technical Journal, 2,* 39-42.

DVB. (2006). Retrieved from http://www.dvb.org/

EWC. (2006). Retrieved from http://www.enhancedwirelessconsortium.org/

Fauconnier, D. (2005, July). HSDPA and HSUPA: UMTS evolution toward higher-bit-rate data. *Nortel Technical Journal, 2,* 13-17.

Flarion Technologies. (2006). Retrieved from http://www.flarion.com/products/flash_ofdm.asp

IEEE 802.21 Task Group. (2006). Retrieved from http://www.ieee802.org/21/

Intel Corporation. (2004a). *Understanding Wi-Fi and WiMAX as metro-access solutions*. Intel Technology White Paper.

Intel Corporation. (2004b). *Orthogonal frequency division multiplexing*. Intel Application Note.

ITU Internet Reports. (2002). *Internet for a mobile generation*. (4th ed.).

MediaFLO. (2005, May 6). *FLO technology brief*. Qualcomm Technology White Paper.

MuniWireless. (2006). Retrieved from http://www.muniwireless.com/

Nortel Networks. (2005). *Introduction to IPTV*. Nortel Networks Position Paper.

OBSAI. (2006). Retrieved from http://www.obsai.org/

Perkins, D. (2005, October 27). *Convergence challenges and solutions for next-generation wireless networking*. Texas Wireless Symposium.

SDR Forum. (2006). Retrieved from www.sdrforum.org/

Tan, C. M., Chin, C. M., & Sim, M. L. (2006). Will different wireless technologies compete or complement one another. Accepted for publication in British Telecom Technology Journal, April 2006.

Teng, R. (2005, January). *Digital multimedia broadcasting in Korea*. In-Stat Report No. IN0502469WHT.

Texas Instruments. (2005). *Digital broadcast TV: Coming soon to a mobile phone near you*. Texas Instruments Technology White Paper.

Thelander, M. W. (2005, October). *The 3G evolution*. CDG White Paper.

Vanjoki, A. (2005, November). Nokia shifts focus to music, photos and TV. *America's Network*, pp. 28-29.

Wi-Fi Alliance. (2006). Retrieved from http://www.wi-fi.org/

WiMAX Forum. (2006). Retrieved from http://www.wimaxforum.org/

KEY TERMS

DMB: A Korean standard for mobile digital broadcast TV, which provides a direct satellite to mobile phone feed.

DVB-H: An open technical specification for bringing digital broadcast services to battery-powered consumer handheld receivers.

FLASH-ODFM: A proprietary wireless communication scheme developed by Flarion (acquired by Qualcomm) that supports high data rates at very low latency over a distributed IP-based mobile network.

FLO: A proprietary standard from Qualcomm designed specifically for multicasting significant volume of rich multimedia content cost effectively to wireless subscribers.

EV-DO: A wireless radio broadband protocol standardised by 3GPP2 for high-speed data transmission.

HSDPA: A packet-based data service in 3GPP system that can achieve data rates of 3-14.4 Mbps (or over 20 Mbps for MIMO systems) over a 5 MHz downlink channel.

HSUPA: An emerging packet-based data service in 3GPP system that can transmit data up to 5.8 Mbps over a 5 MHz uplink channel.

ISDB-T: A standard used in Japan to deliver digital TV services to subscribers.

Mobile-Fi: Nickname for IEEE 802.20 specifications, also known as mobile broadband wireless access.

MIMO: Refers to communication systems that employ multiple antenna elements at both the transmitter and receiver for improved spectral efficiency and channel reliability.

OFDM: A wireless transmission technique that uses many narrowband sub-carriers which are orthogonal to each other in the frequency domain.

QoS: Refers to the capability of a network to provide better service to selected network traffics over various traffic conditions. It is a generic term for measuring and maintaining quality of network characteristics such as error rates, jitter, and latency.

SDR: A communication system that uses software for the modulation and demodulation of radio signals.

UMA: A technology to provide access to cellular networks (2G, 3G) over unlicensed spectrum technologies like Bluetooth and Wi-Fi.

Wireless Fidelity (Wi-Fi): Another name for WLAN based on IEEE 802.11 family of standards.

WiMAX: A trademark refers to broadband metropolitan area network that is based on IEEE 802.16 or HiperMAN standards.

ENDNOTES

[1] Wi-Fi: Wireless Fidelity (Wi-Fi, 2006)

[2] WiMAX: Worldwide Interoperability for Microwave Access (WiMAX, 2006)

[3] HSDPA: High-Speed Downlink Packet Access (Fauconnier, 2005)

[4] HSUPA: High-Speed Uplink Packet Access (Fauconnier, 2005)

[5] EV-DO: Evolution Data Only (or Optimised) (Thelander, 2005)

[6] FLASH-OFDM: Fast Low-latency Access with Seamless Handover - Orthogonal Frequency Division Multiplexing (Flarion, 2006)

[7] DVB-H: Digital Video Broadcasting—Handhelds (DVB, 2006)

[8] DMB: Digital Multimedia Broadcasting (Teng, 2005).

[9] FLO: Forward Link Only (MediaFLO, 2005)

[10] ISDB-T: Integrated Services Digital Broadcasting—Terrestrial (Texas, 2005)

NFC-Capable Mobile Devices for Mobile Payment Services

Stamatis Karnouskos
SAP AG, Germany

INTRODUCTION

An old saying coming from the telecom world states that nothing can be really considered as a service unless you are able to charge for it. As we move towards a service-oriented society, the necessity to pay in real time for a variety of services via different channels anywhere, anytime, in any currency increases. According to Gartner (www.gartner.com), worldwide mobile phone sales totaled 816.6 million units in 2005, a 21% increase from 2004. Due to the high penetration rates of the mobile devices, they pose an interesting candidate for the real-time payment scenarios. Several efforts have already been done (Karnouskos, 2004), but as new technology comes aboard, new capabilities are also brought along. Near Field Communication (NFC) is such a technology, which due to the industry support and its low cost (in comparison with similar ones) may become dominant in short-range communication among a variety of devices, including mobile phones. NFC is well equipped in order to facilitate mobile payments with little interference from the user side.

Mobile Payment

People today use their mobile devices to pay for a variety of mostly intangible goods such as ring tones, games, digital content, and so forth. However existing solutions are confined usually within one service provider and usually consist of local island solutions. The promising trend is to mainly use mobile devices at physical points of sale (POS) and additionally expand the payment capabilities in virtual ones. We consider as mobile payment, any payment where a mobile device is used in order to initiate, activate, and/or confirm this payment (Karnouskos, 2004). A global study by Arthur D. Little Consulting (Taga & Karlsson, 2004) estimates that m-payment transaction revenues will increase from $3.2 billion in 2003 to $37.1 billion in 2008 worldwide. There is evidence of the need for real-time, open, and trusted payment services that can support in a more efficient way the processes evolved in existing electronic and mobile commerce scenarios. Although in the last few years we have witnessed several standardization efforts, the rise and fall of some mobile payment services, promising technologies, ongoing trials, predictions for the future, investments on startup companies, and so forth, there is still no solution that is open, widely accepted, and acknowledged as a clear market leader.

Several reasons exist as to why the mobile payment has not become mainstream such as user friendliness, security, cost, high learning curve for users, lack of the right business models, lack of advanced technology in devices and mobile networks, non-existent cooperation among the key players, and so on. Implementation of mobile payment services is more complex than originally thought and to provide a viable solution has been proven challenging both at the technology and business level. NFC could be one of the enablers that can lead us into effectively tackling some of the issues that have hindered other mobile payment approaches; therefore it is interesting to look at its capabilities and the context of its usage in mobile payment scenarios.

Near Field Communication

NFC is an interface technology for exchanging data between electronic devices. It represents the second generation of the proximity contactless technology, which supports peer-to-peer communication and enables access to services, anytime, anywhere, with any type of NFC-enabled stationary or mobile device. As NFC-compliant devices are brought close together, they detect each other and begin to communicate. This is done at small distances of about 10 cm (4 inches). NFC is based on RF technology at 13.56 MHz, is standardized ISO 18092, and is backwards compatible with IS0 14443. The data exchange rate can be up to 424 Kbit/sec (while 1 Mbit/sec is planned). NFC was designed with the goal to be easy and intuitive to use, in order to be successful also among the technology illiterate users.

NFC devices operate in two different modes:

1. **Reader Mode:** This mode allows the communication with other tags which effectively transforms any device to a fully capable tag reader.
2. **Card Emulation Mode:** This enables the device to behave like a tag itself which can be read by other devices in reader mode.

Standardization of NFC is done within the NFC Forum (www.nfc-forum.org), which was launched in 2004 and in

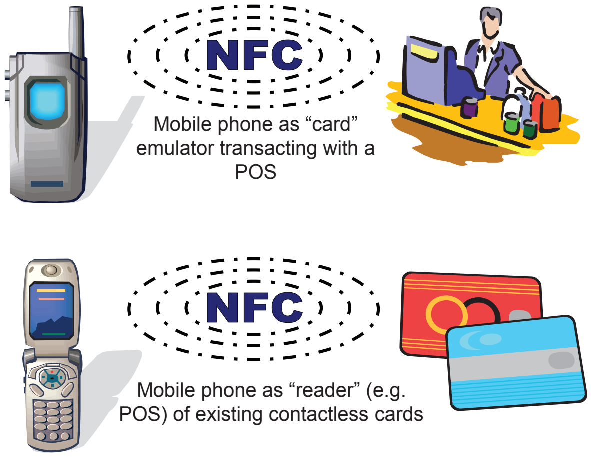
Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the meanwhile has more than 70 members, many of which are key players in their domains and drive the consortium to success. The Nokia 3220 mobile phone (Nokia, 2004) was the first NFC-enabled device that was brought to the market and delivered all the services envisioned by the NFC forum including service discovery, ticketing, and payment. Today other mobile phones also exist such as the Samsung SGH-X700. NFC is compatible with Sony's FeliCa card (http://www.sony.net/Products/felica/) and the broadly established contactless smart card infrastructure based on ISO 14443A, which is used in Philips' MIFARE technology (http://www.semiconductors.philips.com/products/identification/mifare/). This backwards compatibility with the existing infrastructure will ease the introduction of NFC-related services, as not everything has to be done from scratch.

COUPLING NFC WITH MOBILE PAYMENT SERVICES

NFC can be used as a communication protocol for mobile payment applications. In a typical scenario, the user would simply bring in-contact his mobile device with the payment point of sale (POS), and the payment transaction would occur. NFC-enabled devices (e.g., mobile phones) provide an additional security layer since they can transmit encrypted payment information to a POS in a way similar to that used with RFID-enabled credit cards. Furthermore, beyond the existence of an NFC-compliant tag capable of storing a unique ID and transmitting encrypted data at 13.56 MHz using the ISO-19082 air interface protocol, NFC devices also feature a smart card microcontroller. The last can be used as a secure storage for applications and credentials that can be used in payment applications. This allows NFC devices to store data on multiple payment options which provides a flexible base for several business scenarios. Apart from that, the ISO-18092 standard used is also compatible with the ISO-14443A standard, which is currently used by RFID-enabled POS currently installed in several merchants.

Figure 1. The double mode of NFC-enabled devices in payment scenarios


NFC
Mobile phone as "card" emulator transacting with a POS
NFC
Mobile phone as "reader" (e.g. POS) of existing contactless cards

NFC devices can technically operate in two different modes which allows NFC-enabled phones the capability to:

- Fully replace the existing contactless smartcards (when functioning in "card emulation mode"), business and technology wise. All existing business cases that use such cards can now include NFC-enabled mobile phones, which can act as authentication tokens for any transaction.
- Act as a "reader," therefore any reader/POS in the merchant side can be replaced with a mobile phone. Furthermore, the mobility advantage will make it possible to extend existing business cases.
- Make possible new business cases, due to the ability of an NFC-capable device to slip into both modes. For instance in "reader" mode, information can be obtained from a smart advertisement about a concert and the video could be downloaded online from the Internet address specified by the smart tag. The user can pay online and receive the ticket on the mobile phone. Later in "card emulation mode" the user can enter the concert hall by simply waving his mobile (which now has the authentication token stored) from the respective reader.

There are numerous scenarios where NFC-enabled mobile payments can be applied such as gaming, ticketing, purchase of goods, real-time money transfer, and so forth. NFC mobile phones can fully substitute all form of cards as we know them today (credit, debit, prepaid, etc.) and flexibly enable more flexible business models and services to be built.

A number of trials using an NFC mobile phone in order to realize applications that can be hosted under the mobile payment umbrella started within 2005/2006. Most notably:

- In the city of Caen in Normandy, France, trials began in October 2005 (and initially for six months) on NFC-based mobile payments (Caen, 2005). The 200 volunteers in the trial are able to pay with their mobile phone in selected stores (retail), for parking, in tourist sights, and so on. This is the world's first large-scale trial of this emerging technology, and valuable feedback will be obtained from mobile operators, retailers, and consumers. The solution used in this trial incorporates secure, over-the-air (OTA) download of applications on a GSM network and automatically recognizes the appropriate application to launch when an NFC connection is made. The Samsung D500 mobile phone that is used incorporates a Philips smart card chip, enabling users to make payments and use banking applications securely. The process is straightforward: to make a purchase, the customer indicates to the cashier that s/he would like to pay using the phone. The cashier prepares the register to receive payment information via NFC, and then the customer simply waves the phone in front of the terminal.

- In Atlanta's Philips Arena in the United States, season ticketholders had the chance in December 2005 to pay for purchases at concession stands and access mobile content via their Nokia 3220 phones (Philips, 2005).
- In the same spirit, Royal Philips Electronics and Telefonica Moviles España have demonstrated NFC technology at the 3GSM World Congress (www.3gsmworldcongress.com) in Barcelona, by providing 200 selected attendees with NFC-enabled Samsung SGH-X700 mobile phones to be used in a variety of transactions during the event, including secure mobile payment (Philips, 2006). Each phone comes equipped with e-money that can be spent at a specially equipped kiosk at the Philips booth. By using the touch-screen kiosk, users select and pay for their choice of CDs, DVDs, and books using the NFC-enabled phone. After the transaction, the money is deducted from the purse and a message pops up on the phone screen indicating the balance account.
- In Hanau, Germany, the trial "NFC Handy Ticketing" was initiated in April 2005. It enables 200 people who are equipped with the Nokia 3220 mobile phone to use it as an electronic ticket. The customers interact with the electronic legacy ticketing machines that were established in 2002 for RFID contactless tickets, and the data are stored in the mobile phone. The last five trips can be seen anytime, while the NFC function can also be deactivated on demand. The charging is done via the post-payment method, at the end of the month. Controlling the passenger's ticket form is easy, since the controller now equipped with a similar mobile phone simply queries the last ticket data from the passenger's phone (RMV, 2005).

FUTURE TRENDS

The future for NFC looks promising. However there are still several challenges to be mastered before the NFC finds its way into modern application scenarios. NFC brings the promise of gluing the virtual and the real world, and give rise to new innovative services. Coupling it with the mobile phone and a secure environment such as the SIM (Subscriber Identity Module) card, new business cases will emerge which will integrate more mobile phones into our life, eventually even possibly replacing all other tokens that we use today for authentication and payment. Mobile phones could emerge as a global platform and be the common denominator via which service providers will be able to charge for their products in a massive way. However in order for this to be done, new business models need to be developed and new strong partnerships need to be formed. Standardization ensuring interoperability at all levels is crucial when we consider the heterogeneity in hardware and software available in the mobile world. The NFC Forum was founded exactly for this reason and has a promising future.

By coupling NFC with mobile phones and especially the SIM card, network operators (the owners of the SIM) come into an advantageous position. Furthermore, more dynamic and better management of the authentication token can be done since now these tokens can be installed, updated, or revoked via OTA interface. Additionally, if this is coupled with mobile presence information, new security models and risk management mechanisms could emerge. For instance the credit card (stored as authentication token in the mobile phone) could be valid within a geographical area set by its owner. Even more interesting might be cooperative scenarios among such smart tags and context-based services.

NFC can also be used for initial communication, which can eventually result in configuration and further usage of other communication protocols or technologies. Therefore, NFC could be used as a means to initiate mobile payments that can be finalized, such as via instant messaging (Karnouskos, Arimura, Yokoyama, & Csik, 2005). For instance in a taxi-payment scenario, the taxi driver simply touches the taximeter, which registers the amount to be paid in the payment application and provides to the customer all the necessary information (e.g., IM credentials, fare, etc.) for the transaction to continue via an instant messaging platform.

As mentioned NFC-devices can function as "readers" or "card-emulators." However in the middle term, "peer-to-peer" functionality is expected to be added. In that way the reading and/or writing mode would be possible which practically empowers the realization of direct data exchanges (and not via a server) between mobile phones; therefore applications such as file exchange, business cards, and so on would become a reality. For the mobile payment domain, this simply means that payment tokens can flow anonymously from one mobile phone to another. Therefore this form of e-cash (e.g., e-coins), which can be moved among devices via local interaction (no centralized server communication infrastructure needed), has the potential to eventually substitute cash as we know it today.

Security and privacy concerns will have to be fully tackled before NFC-based mobile payment services become mainstream. However, contrary to the traditional payment instruments such as credit cards, mobile devices allow more efficient risk management solutions to be deployed since now, depending on the transaction, password-code or biometric characteristics could be required for high-volume transactions. Furthermore the mobile device can be turned on/off, and the relatively short operational range of NFC at approximately 10 cm constrains (but does not eliminate) possible remote attacks. The security required will be tailored to each service, depending on the risk management and the service's respective business model.

Currently NFC supports ISO 14443A, but ISO 14443B does not. This limits the scenarios where NFC could be used. Mobile phones depend heavily on their batteries. However there are several scenarios where we can use the mobile phone as a simple token and this should not be dependant on its battery status. In other words, in order to cover all possible payment scenarios, we need a technology that does not require that the phone is switched on or its battery charged. ISO 14443B supports such scenarios, but 14443A does not. The extension of the NFC standard to include this capability could expand the use-cases that NFC could play a critical role.

NFC is expected to be complementary to existing protocols (e.g., IrDA, Bluetooth, etc.). Its low cost (around 20¢) is still significantly less, for example, compared to Bluetooth at $5 per item. Due to cost efficiency, as well as the compatibility with existing RFID infrastructure and the large base of smart cards, NFC creates the potential to deliver services effectively anytime, anywhere, and in a variety of channels. Effectively NFC could act as an abstraction layer that would ease the initial communication among devices and bring the vision of easy ubiquitous access to services one step closer to reality. In a service infrastructure, where tangible and intangible goods are offered and can be immediately charged for, mobile payment is expected to be highly integrated and flourish.

CONCLUSION

NFC technology is carefully taking its first steps. Standardization activities have been carefully carried out, and ongoing work within the NFC Forum looks promising. Although there are some prototype mobile phones out there, NFC technology is expected to be integrated in most mobile phones of the near future. "By 2010, we expect that over 50 percent of all mobile handsets will incorporate near field communication chips to enable short-range, easy and secure transactions," points out Erik Michielsen, director at the market analyst firm, ABI Research (www.abiresearch.com). If this will hold true, then mobile payment via NFC has the potential to reach the critical mass rapidly and emerge as an integral part of our future everyday transactions. The first commercial platforms such as the one offered by MobileLime (www.mobilelime.com) are already underway. NFC has learned from previous efforts on new protocol introduction such as Bluetooth, which looked promising but was complex and low on execution. NFC is compatible with existing infrastructure (i.e., Felica and MIFARE) which may give it a significant advantage. Also its nature—such as the short-distance communication and user-friendliness—has an initial positive effect on security and privacy issues, which in any case need to be further investigated. Finally it is pointed out that NFC, amalgamated with mobile payment services, could realize a universal "touch-and-pay" approach anywhere, anytime, in any currency, which in its turn may form the core of more sophisticated business cases. From the market point of view, commercial rollouts could be realized as early as in 2007.

REFERENCES

Caen. (2005). *The NFC trial in Caen.* Retrieved from http://www.caen-ville-nfc.com/

Karnouskos, S. (2004). Mobile payment: A journey through existing procedures and standardization initiatives. *IEEE Communications Surveys & Tutorials, 6*(4). Retrieved from http://www.comsoc.org/livepubs/surveys/public/2004/oct/pdf/KARNOUSKOS.pdf

Karnouskos, S., Arimura, T., Yokoyama, S., & Csik, B. (2005). Instant messaging enabled mobile payments. In A. Salkintzis & N. Passas (Eds.), *Wireless multimedia: Technologies and applications.* New York: John Wiley & Sons.

Nokia. (2005, February). *Nokia announces the world's first NFC enabled mobile product for contactless payment and ticketing.* Retrieved from http://press.nokia.com/PR/200502/979695_5.html

Philips. (2005, December 14). *Industry leaders announce first large-scale near field communication trial in North America.* Retrieved from http://www.semiconductors.philips.com/news/content/file_1209.html

Philips. (2006, February 7). *Philips, Samsung and Telefonica Móviles España demonstrate simplicity of Near Field Communication technology at 3GSM World Congress.* Retrieved from http://www.semiconductors.philips.com/news/content/file_1216.html

RMV. (2005, March). *Weltpremiere in Hanau: RMV startet mit Nokia und Philips pilotprojekt zum handy-ticketing.* Retrieved from http://www.rmvplus.de/getin/NFCHandyTicketing.pdf

Taga, K., & Karlsson, J. (2005, December). *Global m-payment update 2005.* Retrieved from www.adlittle.com

KEY TERMS

Mobile Commerce (M-Commerce): Electronic commerce transactions realized via mobile devices (e.g., mobile phones, PDAs, etc). The term "m-commerce" was coined in the late 1990s during the dot.com boom.

Mobile Device: Any device that can be easily carried around and communicate via mobile/wireless technology. The terms *mobile phone* and *mobile device* are interchangeable in the context of this article.

Mobile Payment: Any payment where a mobile device is used in order to initiate, activate, and/or confirm that this payment can be considered as a mobile payment.

Mobile Ticketing: The realization of a service where virtual tickets are purchased and validated with the help of mobile devices and their authentication capabilities.

Near Field Communication (NFC): A short-range communication technology that can also be used in mobile payment scenarios.

Point of Sale (POS): A location where a transaction occurs. This may be a real POS (e.g., a checkout counter) or a virtual POS (e.g., an e-shop on the Internet).

Subscriber Identity Module (SIM): A smart card that securely stores the key identifying a mobile phone service subscriber, as well as subscription information, preferences, and text messages.

Notification Services for Mobile Scenarios

Michael Decker
University of Karlsruhe, Germany

INTRODUCTION

In this article we introduce the concept of generalized notification services (NSs). NSs are a simple class of services for mobile and wireless terminals, but nevertheless there are many useful services which can be modeled as NSs. Mobile and wireless terminals (MWTs) in our sense are handheld computers with a wireless interface for data communication (e.g., GPRS, UMTS, or WiFi); examples of MWTs are cellular phones, personal digital assistants (PDA), and smart phones. Mobile services are functionalities offered by one or more remote computers (server or back-end systems) to a MWT (client). Because of the mobility of the MWT, at least the first part of the route for the necessary data communication with the server is realized using wireless standards like GPRS, UMTS, or WiFi.

The motivation for giving an exact description of a class of mobile services comes from the fact that in literature there are many descriptions of platforms or technical frameworks for mobile services, but often there is no precise definition of what kind of services are supported by these platforms; these descriptions are more concerned with architectural aspects or how to deal with context information. Also the prevailing paradigm for the realization of mobile services is the Web-like delivery of pull-documents (e.g., iMode) which is not ideal for mobile scenarios, because poor connection quality impairs the user experience (necessity of sending requests again, long waiting times) and push messages are more appropriate for many mobile scenarios.

The remainder of this article is structured as follows: in the next section we cover our understanding of context-awareness. We then explain the basic principle of NSs and argue why they are suitable for mobile scenarios. To show that NSs are a versatile class of mobile services, we mention several examples from different areas of application. Before we summarize, we sketch a protocol for the implementation of NSs.

CONTEXT-AWARENESS OF MOBILE SERVICES

Context-awareness of mobile services (and applications) is an important concept of mobile computing. Context is defined as: "information that can be used to characterize the situation of an entity" and "is considered relevant to the interaction between a user and application" (Dey, 2001). Since almost everything can be considered as an entity at some level of abstraction, we focus more on the purpose of context: to support a user when interacting with a service, for example, displaying information relevant to his or her current situation, or reducing the amount of data to enter manually. The context information has to be available in an explicit form during the runtime of the system. As discussed below in more detail, mobile services have several restrictions with regard to ergonomic aspects, so context information is crucial for the user experience of mobile services.

For the description of NSs, we need to distinguish how critical a given type of context information is with regard to data protection. Context information that is critical for the user's privacy is denoted as "personal context," otherwise as "public context." An example for personal context would be the current location of a user, because people do not like the idea of having their position tracked. "Time" or "weather" (e.g., a mobile service should not recommend outdoor activities when it is raining all day long) are public context parameters because they are not person-related data (but to interpret them, it could be necessary to have personal context information, e.g., the user's location to find out the correct time zone or weather).

BASIC PRINCIPLE

The basic idea behind notification services is that users in mobile scenarios do not want to browse for the information they need in longish sessions and thus push mode for information delivery is preferable. A notification service sends push messages to users based upon an initial configuration and public as well as private context information. Munson and Gupta (2002) already mentioned the idea of generalized notification services, but they considered only location information as a context parameter and their discussion of details concentrated on how to provide location information for a large number of clients.

Our definition of NS demands a special client application on the MWT and a machine-readable service description file for each type of NS. Using the service description the client application can guide the user through the configuration of the NS without any network interaction. The description file also specifies which personal context parameters the client application has to provide (e.g., current location of

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the user retrieved from a GPS module, profile information, battery level).

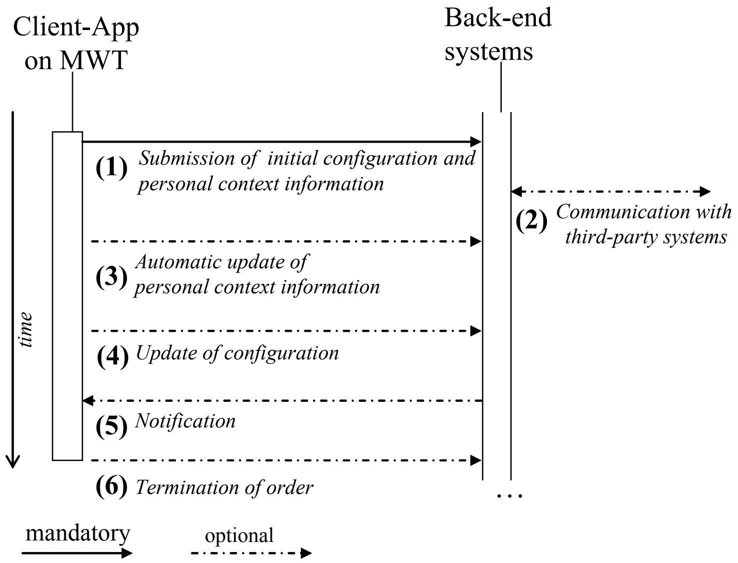
A configured instance of an NS is denoted as "order." The order is submitted to the server where the actual business logic of the NS is running. If certain events occur, the server will send push messages (notification) to the MWT (e.g., SMS/MMS, e-mail, or a client-specific channel based on TCP/IP connection). The content of the push message might lead to further interaction (e.g., the message might contain a link to a WAP-document with further information), but this is beyond the scope of the NS. Depending on the type of order, it might be allowed to reconfigure the order using the client application; also the client application might submit updates of the personal context information (e.g., new position of user) to the server.

To illustrate this concept we consider the example of location-based advertising (Kölmel & Alexakis, 2002). In this scenario a mobile service sends advertisement messages to a user's MWT if the user approaches stores with offers matching his wish list (configuration) and his profile. If implemented as NS, the configuration would describe which offers the user is interested in (e.g., clothes, entertainment, restaurants, etc.); according to his or her current position and profile information (personal context information), the user could receive notifications (advertising messages), for example, about a restaurant not far away from his or her current position.

In Figure 1 we illustrate the messages exchanged between the client on the MWT and the back-end systems of a generic order as a UML sequence diagram (Fowler & Scott, 2000). Steps 2-6 might occur several times, the order may vary:

1. **Submission of Initial Configuration and Personal Context Information:** This step is the only mandatory one. If the NS needs personal context parameters, these are automatically filled in by the client application (e.g., current location of the user or profile information).
2. **Communication with Third-Party Systems:** The back-end systems of the NSs may communicate with third-party systems (e.g., databases to query data or public context information). If the third-party systems proactively send information to the back-end systems (which do not query if they need information), the NS is following the publish/subscribe paradigm which is considered suitable for mobile services (Huang & Garcia-Molina, 2004).
3. **Automatic Update of Personal Context:** If during the lifetime of an order a relevant personal context parameter should change (e.g., location of user, battery level), the client application automatically sends an update to the back-end system.
4. **Update of Configuration:** The user can change the configuration of the order at any time; this might not be reasonable for all kinds of NSs.
5. **Notification:** If the business logic of the NS on the back-end systems detects an event, it dispatches a push message to the MWT. A push message is a message sent to a user without being perceived as being directly requested. There might be multiple notifications or not a single one depending on the type of service.
6. **Termination of Order:** Orders might terminate themselves (e.g., based on configuration parameter "expiry date") or be terminated by the user manually.

Figure 1. Sequence diagram for a generic NS


Client-App on MWT
Back-end systems
time
(1) Submission of initial configuration and personal context information
(2) Communication with third-party systems
(3) Automatic update of personal context information
(4) Update of configuration
(5) Notification
(6) Termination of order
...
mandatory
optional

An important feature of an NS is that both directions of message exchange (upstream: from mobile client to back-end; downstream: notifications from back-end to client) are not coupled, as are services based on document pulls. This feature makes it easy to hide temporary connection drop-outs, but seems to limit the application examples to simple alert services at first glance.

SUITABILITY FOR MOBILE SCENARIOS

Before we can discuss why NSs are particular appropriate for mobile scenarios, we must look at the specific limitations of MWTs (Forman & Zahorja, 1994):

- **Limited Connectivity:** When designing mobile services, possible connection drop-outs, limited capacity, and high latency should be taken into consideration.
- **Limited Ergonomics:** Due to their limited size, MWTs only have a small display, and it is cumbersome to enter data because there is no full keyboard or a "mouse." So a mobile service should not require a lot of data input from the user.
- **Limited Resources:** MWTs possess only limited energy supply, so mobile services cannot perform extensive computations and transfer data over the wireless interface all the time. While currently the available CPU-power and memory on MWTs are still growing, we cannot assume a significant growth of battery capacity in the near future.
- **Privacy Concerns:** Since MWTs can be located by the MNO—he knows always the position(s) of the base(s) station used by the MWT—people are afraid of being "tracked" (Junglas & Spitzmüller, 2005). People also store personal data on their MWTs and do not want to share that data with everyone.

NSs give consideration to these four characteristics:

1. During the configuration, no interaction with the remote server is required; should a connectivity failure occur, the client application will try to resubmit the order later. When using mobile services based on pull documents, the user experience will be impaired by poor connection quality (necessity to manually resubmit request, significant waiting times between pages).
2. NSs require only a minimum of user input; after the initial configuration it usually is sufficient to read incoming notification messages.
3. Due to limited resources, the NS approach does not require extensive computations on the MWT since the actual business logic of the services is operated on a stationary server. Most NSs generate a data volume of a few KBytes per day, even if there are frequent updates of the personal context parameters "location." Only in cases where the notification messages contain extensive multimedia elements (movie sequences, high-quality sound clips) is there a significant amount of data traffic generated by NSs; this might be the case for m-advertising implemented as NS since advertising often uses multimedia elements.
4. Regarding privacy concerns, in the section "NS and Privacy" we show that it is simple to implement NSs in a way to ensure certain requirements with regard to end user privacy.

EXAMPLE SERVICES

We now give examples for mobile services that could be modeled as NSs; for each example we state the required configuration and context parameters, as well as the content of the notification. The example services are not new services; the novelty is their consideration as generalized notification services.

Tourist Guide

This service is for visitors to new places; as they wander around, they receive notification messages with explanations concerning the sights and places in their nearer surrounding. The required personal context parameters are location and profile (e.g., consider age to deliver explanations adequate for children). The time as public context parameter could be considered to mention if the facilities are open or not. See also Sampat, Kumar, Prakash, and McCrickard (2005) for a discussion of a notification system in this usage scenario.

Alert Services

Alert services send notification to users when a time-critical external event was detected, for example, significant changes of stock quotes, emergencies, news, announcement of exam results, or inventory of a certain item got below the minimum threshold (Adya, Bahl, & Qiu, 2002). The configuration describes the event the user is interested in and might also contain a pass phrase, because not everybody is allowed to know about the inventory of a certain firm or the exam results of a certain student. There are even examples of location-aware alert-services, which warn users (e.g., outdoor sportsmen, homeowners, motorists) when a dangerous weather situation is about to appear at their current location. The notifications provide the time-critical information.

Online-Banking Support

To secure online banking in the Internet against so called 'phishing' attacks—the user erroneously enters the secret transaction number (TAN) into the Web site of the attacker—a bank could also utilize the MWT as additional credential for the transaction. This is a form of two-factor authentication (2FA) since the user needs two things—this TAN and his MWT—to prove that he or she is authorized to perform the transaction (Wüest, 2005). The Web site of the bank displays a challenge code after the user has entered the transaction details, which are amount of money, recipient, and TAN. This challenge code is the configuration parameter for an NS, and the network authentication is a context parameter; the notification is a summary of the transaction details and a response code. To complete the transaction the response code from the MWT must be entered into the online-banking Web site. The idea behind this authentication method is that it is very unlikely that the attacker compromises not only the user's PC, but also his MWT. This is an example for an NS where no updates of the configuration or context parameters are allowed. The instance is also terminated automatically after dispatching the response code.

Virtual Memo

Users can deposit "virtual memos" or "virtual graffiti" at their current position (Persson, Espinoza, Fagerberg, Sandin, & Cöster, 2002) with messages like "don't visit the museum in this street." The content of the messages is specified in the configuration (with optional expiry date). If other users of this service approach the location of the memo, they will receive notifications with the content of the deposited memo.

Community Support

Groups of people agree on a group-identifier and an optional pass phrase (configuration). A user now can configure a "Buddy-Finder" NS by providing the identifier and the pass phrase. He or she will receive a notification if another user of that group approaches his or her current situation. If instead of a group-identifier with pass phrase, the profile of people is used, we obtain a "blind-date" finder with notifications like "someone matching your profile is around less than 100 meters."

Mobile Payment

Mobile payment (m-payment) is the usage of a MWT by at least the payer to make a payment (Pousttchi, 2004), for example, payment at a point-of-sale like a vending machine or a supermarket. Configuration parameters are the amount to pay, a pass phrase for authentification, and an identifier of the recipient (e.g., alphanumerical code displayed). The notification would be a confirmation of the payment.

Call-a-Taxi

This location-aware service (also termed "call-an-ambulance" or "call-a-squad-car") calls a taxi to the current position of the user at each invocation of the service; the notification in this case is the confirmation of the order with an optional estimation of the time to wait. This is helpful because tourists or people in emergencies often do not know their location with the accuracy required to be found by a motorist.

Mobile Ticketing

The idea behind m-ticketing is to use MWTs for the distribution and presentation of electronic tickets. A ticket is something that certifies that the owner of the ticket has the right to claim a certain service like public transport (e.g., Böhm, Murtz, Sommer, & Wermuth, 2005) or permission to entertainment events. To hinder fraud, electronic tickets usually have some kind of digital signature. If implemented as NS, a mobile ticketing service sends the tickets as notifications. The configuration of the NS may specify the category of the ticket to buy, for example, seat category for entertainment events or number of zones for public transport tickets.

NOTIFICATION SERVICES AND PRIVACY

As discussed above users have concerns with regard to privacy when using MWT. We make the assumption that there is a trusted zone between MWT and the actual service provider (SP). Using public key encryption (Schneier, 1996), it is simple to state a protocol for the implementation of NS that guarantees the following two properties:

- The trusted zone cannot learn more than necessary about the details of the orders and thus compiles no profile about the fields of interest of the user.
- The SP cannot learn more about the user's identity than necessary for the execution of the service. In particular he or she cannot learn his or her end address, which avoids the danger of unsolicited push messages ("spamming").

Although a single trusted zone between MWT and SP represents a single point of attack, this model is a realistic one and is used by several authors (Bettini, Wang, & Jojodia, 2005). We interpret the trusted zone as the core network of the mobile network operator (MNO), because this part of the network is no public network and a user has to trust the MNO to a certain degree because he or she can track the position and knows when (but not necessarily what) data is transmitted. The actual business logic of the NS is operated by the SP; we assume that MNOs do not provide mobile services because their core business is the provision of wireless communication capacity. It is also assumed that the standard for wireless data transmission provides encryption so the data exchange between MWT and MNO can be considered as secure.

We denote the result of a message *m* encrypted with the public key of participant *a* by $[m]_a$. *X* is the configuration of an order along with the personal context information, SID is a randomly chosen identifier, and *ad* the user's end address. If the MNO has to provide context information, this is mentioned in *Z* (e.g., location information if the MWT has no GPS receiver attached) and thus the MNO has to provide the location information. The responsible service provider is specified by *i*, the public key *SP(i)* is included in the service description file. The client application on the MWT sends an order (or update of an order) in the following form to the MNO:

$$i, Z, [X, SID, [ad]_{MNO}]_{SP(i)}$$

The MNO (he or she cannot decrypt *X*) evaluates *Z* (fills in requested values) and forwards this message to the responsible SP. The SP obtains *X* and *SID* by decryption with his or her private key and creates an order instance (if *SID* is hitherto unknown) or updates the instance with identifier *SID*. If an order instance has to dispatch a notification message *Y* to the end user, the SP sends $Y,[ad]_{MNO}$ to the MNO. He or she can decipher *ad* and thus knows to whom he or she must send the notification message *Y*. Not all notification channels like SMS/MMS or WAP-Push allow encryption, so the MNO who has to dispatch the notification will see *Y* in clear text and maybe can draw conclusions about the order. If the notification channel allows symmetric encryption, *Y* can be decrypted with a symmetric algorithm and a key derived from *SID* since the MNO cannot read *SID*.

SUMMARY

We introduced generalized notification services (NSs), a class of services suitable for scenarios with mobile and wireless terminals. Several examples from different fields of applications were mentioned to show that despite their simplicity, NSs are versatile. A protocol was sketched which helps to guarantee privacy.

Based on the precise description of NSs given in this article, a software framework could be designed. Frameworks represent generic applications of a specific domain and help to develop applications of that domain with reduced effort.

REFERENCES

Adya, A., Bahl, P., & Qiu, L. (2002). Characterizing alert and browser services of mobile clients. *Proceedings of the USENIX Annual Technical Conference* (pp. 343-356), Monterey, CA.

Bettini, C., Wang, X. S., & Jojodia, S. (2005). Protecting privacy against location-based personal identification. *Proceedings of the Conference on Secure Data Management* (pp. 185-199). Trondheim, Norway.

Böhm, A., Murtz, B., Sommer, C., & Wermuth, M. (2005). Location-based ticketing in public transport. *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems* (pp. 837-840). Vienna, Austria.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Forman, G. H., & Zahorja, J. (1994). The challenges of mobile computing. *IEEE Computer, 27*(4), 38-47.

Fowler, M., & Scott, K. (2000). *UML distilled: A brief guide to the Standard Object Modeling Language.* Boston: Addison-Wesley.

Huang, Y., & Garcia-Molina, H. (2004). Publish/subscribe in a mobile environment. *Wireless Networks—Special Issue: Pervasive Computing & Communications, 10*(6), 643-652.

Junglas, I. A., & Spitzmüller, C. (2005). A research model for studying privacy concerns pertaining to location-based services. *Proceedings of the 38th Hawaii International Conference on System Science* (p. 180b).

Kölmel, B., & Alexakis, S. (2002). Location based advertising. *Proceedings of the 1st International Conference on Mobile Business (ICMB)*, Athens, Greece.

Munson, J. P, & Gupta, V.K. (2002). Location-based notification as a general-purpose service. *Proceedings of the 2nd International Workshop on Mobile Commerce* (pp. 40-44). New York.

Persson, P., Espinoza, F., Fagerberg, P., Sandin, A., & Cöster, R. (2002). GeoNotes: A location-based information system for public spaces. In *Designing information spaces: The social navigation approach* (pp. 151-173). London: Springer.

Pousttchi, K. (2004). An analysis of the mobile payment problem in Europe. In *Mobile business systems, mobile and collaborative business, techniques and applications for mobile commerce (TAMoCO)* (pp. 260-268). Essen, Germany.

Sampat, M., Kumar, A., Prakash, A., & McCrickard, D. S. (2005). Increasing understanding of a new environment using location-based notification systems. *Proceedings of 11th International Conference on Human-Computer Interaction,* Las Vegas, NV.

Schneier, B. (1996). *Applied cryptography* (2nd ed.). New York: John Wiley & Sons.

Wüest, C. (2005). Phishing in the middle of the stream—today's threats to online banking. *Proceedings of the 8th Association of Anti-Virus Asia Researchers Conference (AVAR 2005)*, Tianjin, China.

KEY TERMS

Context: Information deliberately used to support a user during his interaction with an application or service; must be available at runtime in explicit form.

Mobile Payment: Procedure where at least the payer uses a mobile and wireless terminal to process a payment.

Mobile Service: A set of functionalities offered by a fixed computer (server) to mobile and wireless terminals (client), whereas the mobile client uses wireless data transmission to communicate with the server.

Notification Services: Class of mobile services where the user receives notification messages in push mode based on a manual configuration and provision of personal context information.

Order: Configured instance of notification services.

Push Message: A message sent to a user which is perceived as not being directly requested. The opposite is "pull message." Push messages are essential for notifications concerning time-critical events, but also bear the risk of being annoying.

Sequence Diagram: One of the diagrams of the Unified Modeling Language (UML) which depicts the temporal order of the messages exchanged between a system's components.

An Ontology-Based Approach for Mobile Agents' Context Awareness

Nejla Amara-Hachmi
University of Paris 13, France

Amal El Fallah-Seghrouchni
University of Paris 6, France

INTRODUCTION

Mobile agents are software agents that can travel among computers under their own control. They can be applied with significant advantages in many domains like network management, information filtering, and electronic commerce. They are especially attractive for performing complex, tedious, or repetitive tasks in open and dynamic systems (Fugetta, Picco, & Vigna, 1998).

Nowadays, mobile agents' applications have to operate within environments having continuously changing execution conditions that are not easily predictable. They have to dynamically adapt to changes in their environment resulting from others' activities and resources variation. To survive, mobile agents have to be aware of their execution context and to have flexible architecture enabling them to envisage an adaptation easily. To do so, it is necessary to have an architecture with two clearly decoupled parts: the mobile agent functional compounds and those ensuring the context handling.

In a previous work (Amara-Hachmi & El Fallah-Seghrouchni, 2004), we proposed a component-based generic adaptive mobile agent (GAMA) architecture that exhibits a minimal mobile agent behavior. In this article, we will focus on GAMAs' awareness of their execution context. Thus, we propose a formal model of context to be used in a semantic approach for checking agents' compatibility with new execution contexts.

The remainder of this article starts with definitions of context and context-awareness for GAMA mobile agents. We then introduce our context model and detail the proposed approach, before concluding the article.

MOBILE AGENTS: CONTEXT AND CONTEXT-AWARENESS

Definitions

As stated in Dey (2001), context consists of "any information that can be used to characterize the situation of an entity" where an entity is "is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." This definition stresses the relation between the system, the context, and the user. In our work, by context we mean information about the current execution environment of the mobile agent, and we focus rather on how the context influences the agent behavior when executing its assigned tasks. This can be illustrated by a failure of the agent execution process if it moves to a host where the context attributes do not fit the agent's requirements. According to these considerations, we propose a definition of the GAMA agents' context-awareness as follows:

Definition 1. A mobile agent GAMA can be context-aware if it is able to detect contextual situations that affect its behavior aiming to achieve its tasks.

According to these definitions, developing context-aware mobile agents requires facilities for recognizing and representing context in order to enable agents to reason on it and make decisions about their execution process (adapt, inform the user, etc.). For GAMA agents, sensing and structuring context information is performed at the hosting platforms. Every interaction between agents and the operating systems (of computers and the network) as well as those between agents and their owners (human users) are achieved by the platform.

The Context Elements

To describe the agents' context, we identify three specific context levels: physical context, social context, and user context. Physical context refers to physical devices the mobile agent is running on, for example, the host's processing power and the network bandwidth. This information is sensed using different probes installed on each operating system. Social context refers to the local multi-agent system (deployed on the platform) with whom the incoming agent has to interact. For instance, information about this context includes the local interaction and coordination protocols. User context refers to the agent's owner preferences, such as restrictions about the exchanged files' size or type, the display quality, and so forth. The user expresses his or her preferences through the platform graphical user interface.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

How To Achieve Context-Awareness

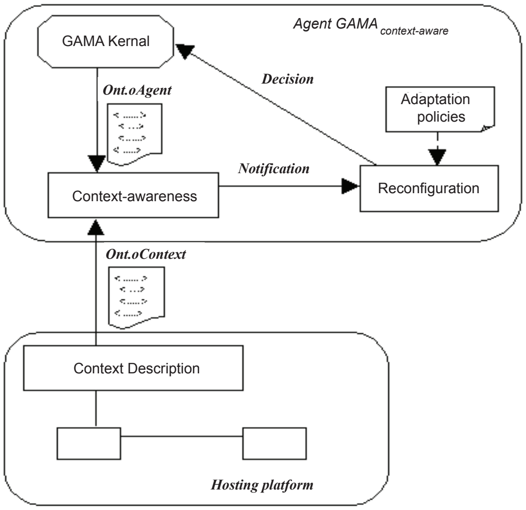
Developing context-aware mobile agents needs capacities to:

1. Capture of the contextual parameters by the means of some physical sensors, graphic interfaces, and the deployment platform. Currently, this step is outside the focus of our work.
2. Model the context by providing formal models of the rough contextual parameters in order to be available for use by the agents.
3. Reason on the context model in order to be aware of it. The reasoning will enable the mobile agent to check if it is able, using its current configuration, to achieve its goals in the new contextual situation to which it moves.

The result of this process is a notification that enables the agent to make a decision about the behavior to adopt in the new context: adapt itself to continue its execution, or alert the user if the adaptation is not possible. Thus, the agent needs not only a model of its context entities, but also a comparable model of its own components. Mapping these models onto one another allows checking their compatibility degrees.

Assuming these requirements, we propose to extend the GAMA architecture with a new component called 'context-awareness'. This component processes the context model provided by the component 'context description" of the platform, and the agent model provided by the component agent profile of the agent. The GAMA architecture is baptized from now up to $\text{GAMA}_{\text{context-aware}}$ (see Figure 1).

Figure 1. From GAMA to $\text{GAMA}_{\text{context-aware}}$



ONTOLOGY-BASED CONTEXT MODEL

To model context, a number of formal and informal approaches exist. We quote specially attribute-value tuples (Dey, Salber, & Abowd, 2001), entity-relationship models (Henricksen, Indulska, & Rakotonirainy, 2002), and first-order predicates (Ranganathan & Campbell, 2003). These representations have the advantage of addressing a certain level of context reasoning, but they offer a weak support for knowledge sharing and are deprived of semantic.

In our work, we need a context representation that: (1) can define common vocabularies to be shared by different agents, and (2) provides a context description at a semantic level in order to enable agents to reason on it. Ontologies seem to be a reasonable solution that meets these requirements. Thus, we propose to develop two distinct ontologies: the first models the context entities *OntoContext,* the second represents the agent components *OntoContext.* These application ontologies are built using the Web Ontology Language OWL, the latest standard of the Web-Ontology Working Group.

A Uniform Frame for Ontologies

If we consider the scenario of a coordinating GAMA mobile agent, an example of contextual attributes that influence its execution process is the type of the negotiation protocol used by the multi-agent system hosted at the visited platforms. The agent must be able to check if it uses the same type of protocol by comparing the parts of two ontologies, *OntoAgent* and *OntoContext,* describing the used protocols.

Thus, in these ontologies, the components providing the negotiation features (of the agent and the hosting platform) must be modeled in a uniform way in order to be able to test their compatibility. Indeed, we consider that compatibility between the agent and its context is ensured whenever the agent's required services (respectively, provided) are provided (respectively, required) by its context. That is why we need a generic and uniform representation of the different agent components and the context entities. Genericity is motivated by the need to model all the entities whatever their origin, and the uniformity aims at facilitating the process of mapping the ontologies.

Representing Entities

Within our working ontologies, we propose to model each component of the agent and each entity of its context by a concept (class, in OWL). These concepts have properties (in OWL, DatatypeProperty) that represent their general characteristics such as the identifier and whether the entity is replaceable or not (in the case of the agent components). Thus, we propose to define a generic concept, #ENTITY, having common properties of the agent components and the context entities. Thereafter, each concept modeling a component in *OntoAgent* or an entity in *OntoContext* will be a subclass of the concept # ENTITY.

An additional requirement mentioned when we defined the GAMA agent's context-awareness was the need to verify if the agent shares with its context common knowledge of the world at a semantic level. The idea is to match only parts of the two ontologies that are semantically close. Therefore, we propose to associate to each entity in the application ontologies a reference to a kind of "mediator" that represents information characterizing a particular domain. This mediator, represented by a taxonomy, plays the role of a domain ontology, and it is used as reference for the semantics of the application ontologies concepts. Taxonomy is a set of hierarchical concepts connected with is-a relationships. It can be represented by an acyclic directed graph, where the concepts are represented by the nodes of the graph connected by directed *is-a* edges. In our work, we use a taxonomy of reference per domain such as the FIPA coordination protocols taxonomy. Considering this, we assign to the concept #ENTITY a new datatype property Taxo_ref with an xsd:string range.

Finally, the generic concept # ENTITY will have the following datatype properties:

- **E_Id (range** xsd:string**):** Indicates the identifier of the modeled entity;
- **Is_replaceable (range** xsd:Boolean**):** Allows, when an adaptation of the agent structure is considered, to know if the modeled component can be replaced by another; and
- **Taxo_ref (range** xsd:string**):** Indicates the taxonomy to which the concept refers.

Representing Attributes

In addition to these properties, each component in the agent (or entity of the context) has a set of provided and required services. In the case of *OntoContext*, the values of these services represent the contextual attributes characterizing the agent execution context. The services of agent components correspond to their interfaces and can be classified as functional and operational. Functional services represent those produced and consumed by the component, while operational services are necessary to connect the component to its hosting architecture. This distinction of services is fundamental to the matching process aiming to test the agent and context compatibility.

To model these services, we define another generic concept #ATTRIBUTE having the following datatype properties:

- **At_Id (range** xsd:string**):** indicates the identifier of the service;
- **At_Type (range** Any**):** allows at the instantiation consideration of various datatypes such as string, date, integer, etc.;
- **Is_Prov (range** xsd:bool**):** indicates if the service is provided (at instantiation will have as value, true) or is required (value to false); and
- **Is_Func (range** xsd:bool**):** indicates if the attribute models a functional service (its value is true) or operational (its value is false).

Each service in *OntoAgent* and *OntoContext* will then be represented by a concept that is a subclass of the generic concept #ATTRIBUTE. These concepts modeling the services are connected to the concepts representing entities via ObjectProperties.

OntoAgent and *OntoContext*

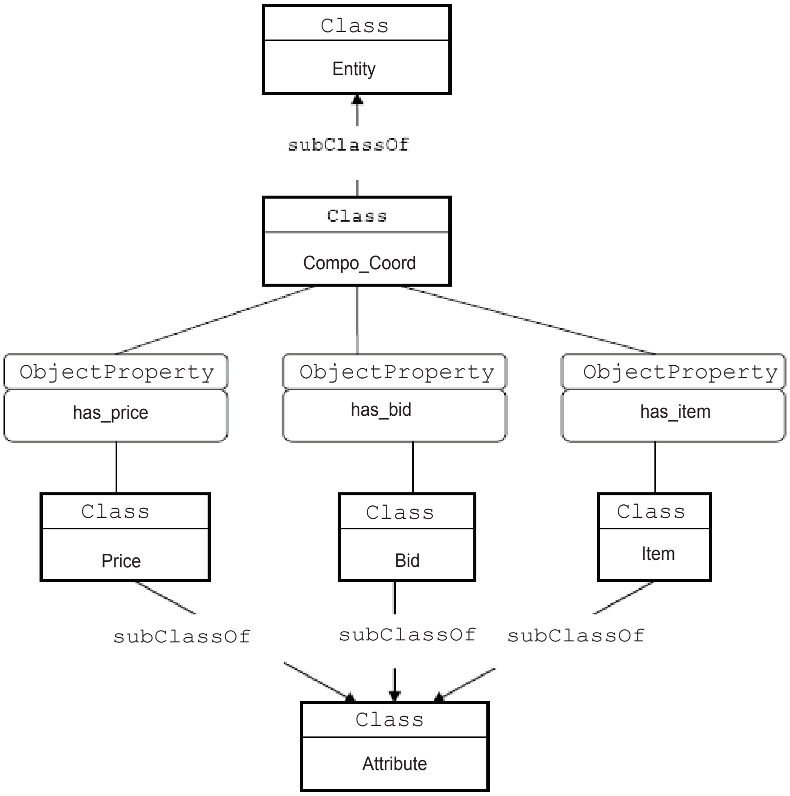
OntoContext is the context ontology. Its class definition begins with the class Context that represents an abstraction of the agent execution context. This class is on the top of the hierarchy of subclasses characterizing the agent context. *OntoContext* defines three subclasses of the Context class—PhysicalContext, SocialContext, and UserContext—in order to represent the several context aspects. Each one of these subclasses is related to a set of classes that characterize the correspondent context features. For example, the PhysicalContext class is related to the classes ComputerMemory, ComputerCPU, and NetworkBandwidth to express that the system level of the agent context can be described by the features of the current host, such as its memory size and CPU power, and the bandwidth of the used network.

OntoAgent is the agent profile ontology. Since a GAMA agent is built by assembling software components, its profile can be described by an ontology including descriptions of all its components. The top-level class of the *OntoAgent* ontology is *AgentProfile*, which represents an abstraction of information about the agent structure. It has several subclasses describing the different functional components of the agent and properties representing relations between theses classes. For instance, the coordination component of a GAMA agent can be modeled using the introduced generic concepts as depicted in Figure 2.

SEMANTIC WEB TECHNIQUES FOR GAMA AGENTS' CONTEXT-AWARENESS

The third step in the GAMA agents' context-awareness process is the reasoning upon the context and the agent ontologies in order to check their compatibility on each visited host. The main rationale behind this reasoning is

Figure 2. Model of the agent coordination component



to first, map parts of *OntoAgent* and *OntoContext* that are semantically related, and check after that the compatibility of services required and advertised by the mobile agent and its context. To do so, we propose a two-step semantic approach based on a lightweight method for mapping ontologies and a matching algorithm of the mapped concepts.

In this section, we describe first the proposed semantic approach. Then we present the mapping method and the matching process.

Method

In this work, we assume that for being aware of its execution context, a GAMA agent must be able to answer the following questions: (1) Does it share with its context a common knowledge of the World at a semantic level? (2) Is it able to provide adequate services to interact correctly with its context? (3) Is it possible to meet, on the new visited context, the suitable services required for its execution? In a nutshell, a GAMA agent is expected to be able to "understand" new contexts that it visits and decide whether they are adapted for its execution.

To satisfy these desiderata, we propose a reasoning process to be held by the mobile agent component, called "context-awareness." The reasoning is triggered at each migration of the agent and takes as input instances of the ontologies *OntoContext* and *OntoAgent.* The output is a notification sent to another component responsible of making a decision concerning the following agent behavior. This notification aims to specify if the agent is able to correctly continue its execution on the host on which it arrives and otherwise to detail causes of their incompatibility. In this last case, the agent can envisage a dynamic reconfiguration of its structure in order to adapt to the new context.

To achieve the stated goal, we introduce a canonical reasoning process that uses the working ontologies to produce the desired notification. Processing ontologies consists first in mapping concepts' modeling entities supposed to have semantic relationships. After that, computed concepts are analyzed in order to detect coherent ones. Finally, the resulting pairs of concepts are matched to check their compatibility.

Mapping Method: LiteMap

To map our ontologies, we are interested in 1:1 relations and will refer to existing mapping approaches that operate according to a common general process (Ehrig & Staab, 2004). Inputs are two ontologies O1 and O2, and the mapping tries to seek for each concept of O1 the corresponding concept in O2 which has the same intended meaning. There are several techniques for mapping ontologies such as terminological, structural, or semantic techniques (Euzenat, Le Bach, Barrasa, Bouquet, De Bo, Dieng, et al., 2004).

The method we propose, LiteMap (for Lite Mapping), accommodates this general mapping process to our problem, and combines semantic and syntactic approaches. However, we use a selection strategy making it possible to simplify and gain efficiency for mapping our ontologies. LiteMap is composed of two macro steps:

1. **Selection of Mapping Candidates:** In many mapping approaches, this step is costly and represents a major source of complexity. This is due to the number of concepts to compare in order to find eventual mapping candidates. Naïve mappings have to compare all concepts of O1 to all concepts of O2. Fortunately, a significant feature of our work is the use of reference taxonomies that ensures a shareable vision of the world to both mobile agents and hosting platforms. This consideration enables reducing considerably the number of concept pairs of *OntoAgent* and *OntoContext* that ought to be investigated. Indeed, we need to compare only concepts that are likely to have close semantics. Moreover, contrarily to usual mapping methods, this step is based on semantic relations holding between concepts and not labels. Our mapping candidates will then be pairs of concepts that have the same reference taxonomy. Let S be the set of mapping candidates:
   $\mathcal{S} = \{(A_i,C_j),\ i \in \{1..n\}$ and $j \in \{1..m\}\}$, where:
    - A_i is a concept of *OntoAgent.*
    - C_j is a concept of *OntoContext.*
    - A_i and C_j have the same **Taxo_ref**.
    - n is the number of *OntoAgent* concepts.
    - m is the number of *OntoContext* concepts.
2. **Computing Relations:** When two concepts of *OntoAgent* and *OntoContext* are candidates for mapping, we then seek to explore the degree of their coherence at a semantic level. For that, we consider three types of semantic relations between two concepts: equivalence, subsumption, and mismatch. We then define semantic coherence as follows.
   Two concepts C1 and C2 are semantically coherent (the relation is symmetrical) if in the same taxonomy of reference we have:
    - C1 ≡ C2 when C1 and C2 are strictly equivalent, or
    - C1 subsumes C2 when C2 is a direct son of C1 in taxonomy (C1 Í C2), or
    - C2 subsumes C1 when C1 is a direct son of C2 in taxonomy (C1 Ê C2).

   We note C1 » C2 and consider that they mismatch in the other cases (C1 ^ C2).

To compute these relations among the set $\mathcal{S}$ of candidate mapping concepts, we use syntactic comparison of concept labels. Indeed, to discover the binding relation R that holds between two mapping candidates concepts from *OntoAgent* and *OntoContext,* we have to compare labels of these concepts to labels of the reference taxonomy concepts. Label comparison is held by measuring linguistical closeness between concept labels. Various measures and metrics can be used to compute string similarities, such as Hamming distance, substring similarity, or N-gram distance. At this stage of work, we use naïve string equality. The output of this step is S-co, a subset of S composed of concept pairs that are semantically coherent:

$$\mathcal{S}\text{-}co = \{(A_i,\ \mathcal{R},\ C_j) \text{ such as } (A_i,\ C_j) \in \mathcal{S}$$

and

$$\mathcal{R} \in \{\equiv,\ \text{Í},\ \text{Ê}\}\}$$

These concepts will be then used in a matching process to check the agent and context compatibility. However, pairs of concepts that mismatch are an immediate cause of incompatibility, thus the implied agent concept is added to the notification set.

Compatibility Test: The Matching Process

The main idea behind the compatibility test between a $\text{GAMA}_{\text{context-aware}}$ agent and its context is to conduct a matching process between their provided and requested services. Indeed, a key observation we already stated is that mobile agents have to respect the subsequent condition to ensure awareness about their context, provide adequate services to interact correctly with the visited host, and meet the suitable services required for its execution. We propose here a matching algorithm that strives to meet these requirements.

Let #E be a concept of *OntoAgent* or *OntoContext.* This concept will be a subclass of the generic concept #ENTITY. Moreover, it is connected (by means of ObjectProperties) to a set of concepts #t1..., #tn that are subclasses of the generic concept #ATTRIBUTE, representing services of the entity modeled by concept #E. We note T(E) the set of these concepts, T(E) $=\{t_1,\ldots,t_p\}$ where p is the number of services of #E. For each service t_i, we define the two following predicates:

- ***is_func*** (t_i), whose value is *true* if t_i is a functional attribute and *false* if t_i is operational.
- ***is_prov*** (t_i), whose value is *true* if t_i is a provided attribute and *false* if t_i is required.

By using these predicates, we can define the following sets:

- Prov (E) = {t_i ∈T(E) | ***is_prov*** (t_i) = *true*, i ∈ {1..p }} that represents the set of provided services of #E.
- Req(E) = {t_i ∈T(E) | ***is_prov*** (t_i) = *false*, i ∈ {1..p }} that represents the set of required services of #E.
- Func (E) = {t_i ∈T(E) | ***is_func*** (t_i) = *true*, i ∈ {1..p }} that represents the set of functional services of #E.
- Opr (E) = {t_i ∈T(E) | ***is_func*** (t_i) = *false*, i ∈ {1..p }} that represents the set of operational services of #E.

Using these sets we can define for *OntoAgent* and *OntoContext*:

- The set of functional provided services: FP(E) = Func (E) ∩ Prov (E).
- The set of functional required services FR(E) = Func (E) ∩ Req(E).

Considering these sets, we can now define semantic compatibility (*S-Compatibility*) between concepts. The key idea underlying this definition is to express that a concept A of *OntoAgent* is semantically compatible with a concept C of *OntoContext* when the services provided by A (respectively C) are used by C (respectively A).

Definition 2. Let A be a concept of *OntoAgent* and C a concept of *OntoContext*. We define the predicate *S-Compatibility* (A,C) having Boolean values as follows:

S-Compatibility(A,C) = *true* if and only if FP(A) Í FR (C) and FR(A) Ê FP(C).

Assuming this, we can precisely note required conditions for compatibility between an agent and its context. A $\text{GAMA}_{\text{context-aware}}$ agent is compatible with the visited context if the corresponding *OntoAgent* and *OntoContext* ensure the subsequent criteria:

For each concept pair (A, C) ∈ *S-co*, such as A ∈ *OntoAgent* and C ∈ *OntoContext,* then *S-Compatibility*(A,C) = *true*.

CONCLUSION

In this article we proposed an approach to address mobile agents' context-awareness using techniques derived from the semantic Web domain. To validate it, we proposed a prototype of a GAMA mobile agent and a hosting platform simulating a pervasive execution environment, implemented with Enterprise Java Beans of Sun. We also have edited the ontologies *OntoContext* and *OntoAgent* using Protégé2000 (http://protege.stanford.edu/), an open source ontology editor and knowledge-base framework. We are actually testing the developed prototype with a use case concerning assistant agents in virtual libraries.

REFERENCES

Amara-Hachmi, N., & ElFallah-Seghrouchni, A. (2005, July). A framework for context-aware mobile agents. *Proceedings of the 4th AmbiAgents Workshop at AAMAS'05* (pp. 67-73). Utrecht.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Dey, A. K., Salber, D., & Abowd, G. D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction (HCI) Journal, 16*(2-4), 97-166.

Ehrig, M., & Staab, S. (2004). QOM: Quick Ontology Mapping. *Proceedings of the 3rd International Semantic Web Conference* (p. 683). Hiroshima, Japan.

Euzenat, J., Le Bach, T., Barrasa, J., Bouquet P., De Bo, J., Dieng, R., et al. (2004). State of the art on ontology alignment. *Knowledge Web Deliverable #D2.2.3.* INRIA, Saint Ismier.

Fuggetta, A., Picco, G. P., & Vigna, G. (1998). Understanding code mobility. *IEEE Transactions on Software Engineering, 24*(5), 342-361.

Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002). Modeling context information in pervasive computing systems. *Proceedings of Pervasive'02 (LNCS 2414),* (pp. 169-180). Berlin: Springer-Verlag.

Ranganathan, A., & Campbell, R. H. (2003). An infrastructure for context-awareness based on first order logic. *Personal and Ubiquitous Computing, 7*(6), 353-364.

KEY TERMS

Component: Component-based software engineering is a product of the natural evolution of object-oriented languages with assembly capabilities beyond inheritance. While objects are expressed on the language level, components are expressed principally by exploring their public interface and promoting black box reuse.

Context: Consists of any information that can be used to characterize the situation of an entity where an entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.

Mapping: Two ontologies O1 and O2 try to seek for each concept of O1 the corresponding concept in O2 which has the same intended meaning.

Mobile Agent: Software agent that can travel among computers under its own control.

Ontology: A data model that represents a domain and is used to reason about the objects in that domain and the relations between them.

Taxonomy: A hierarchical taxonomy is a tree structure of classifications for a given set of objects.

An Optimal Timer for Push to Talk Controller

Muhammad Tanvir Alam
Bond University, Australia

INTRODUCTION

The push-to-talk over cellular (PoC) application allows point-to-point or point-to multipoint voice communication between mobile network users (Balaz, 2004). The communication is strictly unidirectional, where at any point of time only one of the participants may talk (talker), and all other participants are listeners. In order to get the right to speak, listeners must first push a "talk" button on their mobile terminals. Floor control mechanisms ensure that the "right to speak" is arbitrated correctly between participants. The PoC application may become a highly popular service for the mobile telecommunications market if its responsiveness and voice quality meet end-user expectations. In the autumn of 2003, Ericsson, Motorola, Nokia, and Siemens submitted their jointly defined PoC specifications to the Open Mobile Alliance (OMA, 2005) to facilitate multi-vendor interoperability for push-to-talk products. The specification is based on the Third Generation Partnership Project's (3GPP's) IP Multimedia Subsystem (IMS) architecture (3GPP, 2005); PoC is to bring the first commercial implementations of the IMS architecture into mobile networks. A discussion on strategic actions related to standardization, system architecture, and service diffusion of PoC has been discussed by Vehmas and Luukkainen (2005). An exploratory study of college-age students using two-way PoC cellular radios has been shown by Woodruff and Aoki (2003).

One wireless carrier, Nextel Communications (2002), provides mobile phones with conventional features such as voice telephony and voicemail; the same network and handsets also support a two-way, push-to-talk service called Direct Connect™. This service is very popular, having 10 million subscribers and supporting nearly 50 billion Direct Connect calls in 2001, predominantly for business use (according to a report of Nextel Communications, 2002). Competitors are attempting to introduce similar services based on packet (IP) networking; the top four U.S. carriers have all announced plans for similar services in the very near future, and separate service providers such as fast-mobile (www.fastmobile.com) are also appearing, particularly in Europe.

"Equally important is the fact that push-to-talk is a forerunner to peer-to-peer services over IP, for which IMS provides the capabilities and foundation. PoC is the first commercial application based on IMS" (Northstream, 2004). The driving forces behind the operators' push-to-talk initiatives are the search for new revenue opportunities and finding ways to increase subscriber acquisition and reduce churn. In this article, we depict some of the potential problem areas of a PoC server and provide an analytic model to ameliorate dimensioning of a PoC controller.

BACKGROUND

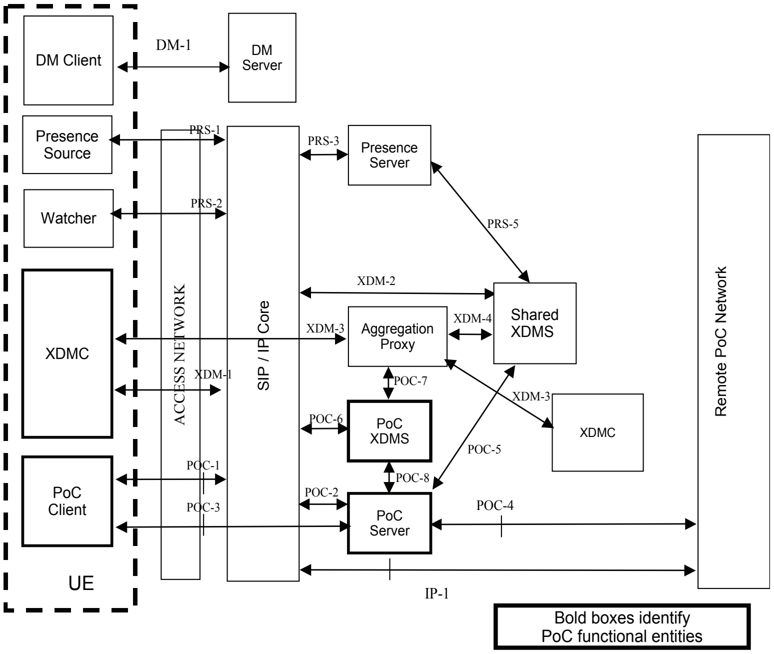
PoC is a warm topic today for the researchers. An architecture for enabling PoC services in 3GPP networks has been furnished by Raktale (2005). Similar work is reported by Parthasarathy (2005). The design of a PoC service operated over a GPRS/UMTS (general packet radio service/universal mobile telecommunications system) network is also depicted by Kim, Balazs, Broek, Kieselmann, and Bohm (2005). The basic architecture of PoC is provided in Figure 1. The common terms related to PoC have been furnished in the Key Terms section. The PoC server implements the application-level network functionality for the PoC service. It performs a controlling PoC function and/or participating PoC function (OMA, 2005). The controlling PoC function and participating PoC function are different roles of the PoC server.

The determination of the PoC server role (controlling PoC function and participating PoC function) takes place during the PoC session setup and lasts for the duration of the whole PoC session. In case of *1-1 PoC session* and *ad-hoc PoC group session,* the PoC server of the inviting *user* performs the controlling PoC function. In case of the *chat PoC group* and *pre-arranged group session,* the PoC server owning/hosting the *group identity* performs the controlling PoC function. The PoC server performing the controlling PoC function has *N* number of SIP (session initiation protocol) sessions and media and talk burst control communication paths in one PoC session, where *N* is number of participants in the PoC session. The PoC server performing the controlling PoC function will have no direct communication to the PoC client for PoC session signaling, but will interact with the PoC client via the PoC server performing the participating functioning for the PoC client.

The PoC server performing the controlling PoC function normally also routes media and media-related signaling such as talk burst control messages to the PoC client via the PoC server performing the participating PoC functioning for the PoC client. However, local policy in the PoC server performing the participating PoC function may allow the PoC server performing the controlling PoC function to have

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. PoC architecture (OMA, 2005)


DM Client
DM-1
DM Server
Presence Source
PRS-1
PRS-3
Presence Server
Watcher
PRS-2
PRS-5
XDMC
XDM-2
XDM-3
Aggregation Proxy
XDM-4
Shared XDMS
XDM-1
ACCESS NETWORK
SIP / IP Core
POC-7
XDM-3
XDMC
POC-6
PoC XDMS
POC-5
PoC Client
POC-1
POC-8
POC-3
POC-2
PoC Server
POC-4
Remote PoC Network
UE
IP-1
Bold boxes identify PoC functional entities

a direct communication path for media and media-related signaling to each PoC client. A PoC server performing the participating PoC function always has a direct communication path with a PoC client and a direct communication path with the PoC server performing the controlling PoC function for PoC session signaling.

The basic challenges that affect the end-to-end service performance for PoC are:

1. network configuration and *dimensioning*;
2. timer settings in terminals and networks;
3. traffic handling priorities used; and
4. service option choices such as early media session establishment.

In this short article, we focus on the issue of *optimized lifetime* of a session provided by the PoC controller. The lifetime of a session is crucial since the server capacity is always fixed. Thus, the lifetime of a session must be set carefully based on the available resources, number of clients willing to talk, network topology, and so forth.

PROPOSED OPTIMAL TIME

Dimensioning and optimizing networks is a mature topic today (Ghaderi & Boutaba, 2006; Vacirca, Vendicits, & Baiocchi, 2006). An optimal design for a multi-rate ATM loss network is provided by Mitra, Morrison, and Ramakrishnan (1996). We use the square root dimensioning method to compute the optimized timeframe for a session of PoC controller. The mechanism is based on the mean service rate of the controller, number of current sessions, and the controller capacity. Let, T_i be the average session time encountered by a job i and C_i the capacity allotted for session i. Also, let λ, λ_i, and μ be the total PoC session arrival rate, arrival rate of specific session PoC i, and mean service rate of PoC sessions respectively.

If the stability condition $\lambda_i < \mu\ C_i$ for $i=1,2,..,N$ holds, then:

$$T_i = \frac{1}{\mu\ C_i - \lambda_i} \quad \forall i = 1,2,..,N \tag{1}$$

Thus the average PoC session time of a PoC messages i is (assuming $\rho_i = \frac{\lambda_i}{\mu}$):

$$T_i = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \left(\frac{1}{\mu C_i - \lambda_i}\right) = \frac{1}{\lambda} \sum_{i=1}^{N} \frac{\rho_i}{C_i - \rho_i} \tag{2}$$

The stability condition now reads $\rho_i < C_i$ for $i=1,..,N$. Therefore, our threshold timeframe problem reduces to:

Optimize: T_i

With respect to: $\{C_i\}_{i=1}^{N}$

Under the constraint: $D = \sum_{i=1}^{N} C_i$

Where, D is the total capacity of the PoC controller.

Applying the Lagrange multiplier technique, we define:

$$f(C_1,...,C_N) = \frac{1}{\lambda}\sum_{i=1}^{N}\frac{\rho_i}{C_i-\rho_i} + \beta\left(\sum_{i=1}^{N} C_i - D\right). \tag{3}$$

Solving the equation $\P f(C_1,..., C_N) / \P C_i = 0$ for every $i=1,2..,N$; we obtain:

$$C_i = \rho_i + \sqrt{\frac{\rho_i}{\lambda\,\beta}} \qquad i=1,2,..,N. \tag{4}$$

From the constraint $\sum_{i=1}^{N} C_i = D$ and equation (4) we get:

$$\frac{1}{\sqrt{\lambda\,\beta}} = \frac{D - \sum_{i=1}^{N}\rho_i}{\sum_{i=1}^{N}\sqrt{\rho_i}} \tag{5}$$

Introducing the value of equation (5) into equation (4) yields:

$$C_i = \rho_i + \frac{\sqrt{\rho_i}}{\sum_{i=1}^{N}\sqrt{\rho_i}}\left(D - \sum_{i=1}^{N}\rho_i\right) \tag{6}$$

Therefore the optimized timeframe of a PoC session *i* becomes:

$$T_{opt} = \frac{\sum_{i=1}^{N}\rho_i}{\lambda\left(D - \sum_{i=1}^{N}\rho_i\right)} \tag{7}$$

Equation (7) may be used by the PoC controller during heavy traffic to optimize a session. The parameters of the equation will be adjusted accordingly. After the optimized lifetime, a session will be disconnected by the PoC controller.

FUTURE TRENDS

The PoC service has yet to undergo further refinement to overcome all of its shortcomings. A new algorithm for efficient dimensioning is required to reduce congestion in the network. If PoC is to carry real-time data over packet switching networks, the Internet standard protocol real-time protocol (RTP) can be used. In that case, low latency and in-sequence delivery need to be guaranteed by some algorithm that would synchronize the clock that timestamps the packets. The PoC signaling in circuit-switched networks is limited to SMS (short messaging service) only. In order to use SMS for push-to-talk, signaling will increase the server load and consume more resources. Thus to use SMS capacity for PoC signaling is expensive, as the traditional use brings direct revenues that are now used for PoC. The cost for this signaling should be compared to the price of sending SMS messages for the end user.

The voice quality, presence functionality, and so forth need to be tested under GSM (global system for mobile communications), GPRS (general packet radio service), and EGPRS (enhanced GPRS) technologies. The different codec facilities—for instance, AMR (adaptive multi-rate voice codec), EFR (enhanced full rate), and so forth—are to be studied both in error-free and error-prone states.

New models could be derived based on the cost incurred by the radio access network to implement PoC service. Some necessary cost items are radio network planning activities, transition network, core network planning, number of application servers used, service integration, and marketing activities. Efficient scheduling is always preferable below the MAC (media access control) layer for the PoC sessions to share the timeslots of the server. The future releases of the OMA should refer to the abovementioned areas at least.

CONCLUSION

This research depicts a potential approach to set the timer for a PoC controller. The mechanism can be used in the similar states for the PoC controller for instance, setting an optimized lifetime for dropping a queued message from the server, setting timer for a session to enter idle state from ready state, and so on. Today, major operators are currently evaluating push-to-talk, and a lot of them have some PoC trials with a number of different vendors. Motorola developed its iDen technology with PoC functionality more than a decade back. Ericsson, Motorola, Siemens, and SonyEricsson announced plans for joint interoperability testing for their PoC products. The major handset manufacturers are all active in the area of push-to-talk. Several agreements between PoC vendors and handset manufacturers have been announced already. Clearly, more commercial PoC launchers are underway, and researchers are up to the task of shaping the service further.

REFERENCES

Ali-Vehmas, T., & Luukkainen, S. (2005). Service diffusion strategies for push to talk over cellular. *Proceedings of the IEEE International Conference on Mobile Business (ICMB)* (pp. 427-433).

Balazs, A. (2004). QoS in wireless networks: Push-to-talk performance over GPRS. *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems* (pp. 182-187).

Ghaderi, M., & Boutaba, R. (2006). Call admission control for voice/data integration in broadband wireless networks. *IEEE Transactions on Mobile Computing, 5*(3), 183-207.

Kim, P., Balazs, A., Broek, E., Kieselmann, G., & Bohm, W. (2005). IMS-based push-to-talk over GPRS/UMTS. *Proceedings of the IEEE Wireless Communications and Networking Conference* (Vol. 4, pp. 2472-2477).

Mitra, D., Morrison, J., & Ramakrishnan, K. (1996). ATM network design and optimization: A multirate loss network framework. *IEEE/ACM Transactions on Networking, 4*(4), 531-543.

Nextel Communications. (2002). Retrieved from http://www.nextel.com/services/directconnect.shtml

Northstream, A. B. (2004). *Overview and comparison of push-to-talk.* Retrieved from www.northstream.se

OMA (Open Mobile Alliance). (2005). *Push to Talk Over Cellular Working Group.* Retrieved from http://www.openmobilealliance.org

Parthasarathy, A. (2005). Push to talk over cellular (PoC) server. *Proceedings of the IEEE International Conference on Networking, Sensing and Control* (pp. 772-776).

Raktale, S. (2005). 3PoC: An architecture for enabling push to talk services in 3GPP networks. *Proceedings of the IEEE International Conference on Personal Wireless Communications (ICPWC)* (pp. 202-206).

3GPP. (2005). *The 3rd Generation Partnership Project, IP Multimedia Subsystem, stage 2.* Retrieved from http://www.3gpp.org

Vacirca, F., Vendicits, A. D., & Baiocchi, A. (2006). Optimal design of hybrid FEC/ARQ schemes for TCP over wireless links with Rayleigh fading. *IEEE Transactions on Mobile Computing, 5*(4), 289-302.

Woodruff, A., & Aoki, P. (2003). How push-to-talk makes talk less pushy. *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 170-179).

KEY TERMS

Application Server: In 3GPP IMS, a functional entity that implements the service logic for SIP (session initiation protocol) sessions.

Controlling PoC Function: Implemented in a PoC Server and provides centralized PoC session handling, which includes RTP media distribution, talk burst control, policy enforcement for participation in group sessions, and the participant information.

Group: A predefined set of PoC users identified by a SIP URI. A PoC client uses the group to establish PoC sessions and to define PoC session access policy.

Home PoC Server: The PoC server owned by service provider that provides PoC service to the user.

On-Demand Session: A PoC session set-up mechanism in which all media parameters are negotiated at PoC session establishment.

Participating PoC Function: Implemented in a PoC Server and provides PoC session handling, which includes policy enforcement for incoming PoC sessions and relays talk burst control messages between the PoC client and the PoC server performing the controlling PoC function. The participating PoC function may also relay RTP media between the PoC client and the PoC server performing the controlling PoC function.

PoC Session Identifier: An identifier in the user plane associated with a PoC session that uniquely distinguishes a particular PoC session from all other PoC sessions, including those that currently exist and those that do not.

Simultaneous PoC Session: When a PoC user is a participant in more then one PoC session simultaneously using the same PoC client.

Talk Burst Control: A control mechanism that arbitrates requests from the PoC clients for the right to send media.

Talk Burst Control Protocol: A protocol for performing talk burst control and defined in these specifications.

User Equipment (UE): A hardware device that supports a PoC client (e.g., a wireless phone).

XML Document Management Client (XDMC): An XML configuration access protocol client which manages XML documents stored in the network (e.g., PoC-specific documents in the PoC XDMS, URI lists used as contact lists in the shared XDMS, etc). Management features include operations such as create, modify, retrieve, and delete. The

XDMC is also able to subscribe to changes made to XML documents stored in the network, such that it will receive notifications when those documents change. The XDMC can be implemented in a UE or fixed terminal.

Optimal Utilisation of Future Wireless Resources

Choong Ming Chin
British Telecommunications (Asian Research Center), Malaysia

Chor Min Tan
British Telecommunications (Asian Research Center), Malaysia

Moh Lim Sim
Multimedia University, Malaysia

INTRODUCTION

Radio resource management (RRM) is one of the most challenging and one of the most important aspects in the provisioning of quality of service (QoS) for wireless communication systems. Conceptually RRM policies, in tandem with network planning and air interface design, determine QoS performance at the individual user level and also at the network level and hence significantly improve system performance. For the past decade, the confluence of users' demand for personal and technological advances in communication systems has led to a phenomenal explosion and deployment of mobile wireless systems, and at the turn of this century we have been preparing ourselves for future mobile communication systems. Technological buzzwords such as fixed-mobile convergence (UMA, 2006), Wi-Fi (Wi-Fi Alliance, 2006), WiMAX (WiMAX Forum, 2006), 3G evolutions (Thelander, 2005), and Mobile-Fi (IEEE 802.20, 2006), together with the advancement of broadband multimedia services, are poised to permeate into every fabric of our society, taking flight into every possible imagination of mobile users. Hence, in such a rapidly evolving and expanding environment, resource management remains a central issue in the very near and distant future.

With the ever increasing size of the wireless mobile community and its demand for high-speed multimedia communications, efficient resource management becomes a paramount importance due to limited resources available, such as spectrum or bandwidth allocation and power availability. The objective of utilising RRM techniques is of course to maximise the performance of network throughput or total resource utilisation while satisfying the wireless users' or service providers' requirements. From the users' point of view, they want to maximise services, such as getting maximum throughput, lowest block and drop rate. On the other hand, from the service providers' viewpoint, they want to maximise revenue; to serve the maximum number of users as possible at a given time. Meanwhile, the service providers would also hope to minimise capital and operational expenditure, for example, deployment of base stations and backhaul links. Hence this article will address the issues of RRM problems in future wireless and mobile networks.

Given the fact that both broadband multimedia services and wireless services are the twin engines driving the future growth of telecommunication industry, current RRM techniques are certainly inadequate to face future demands from the wireless community. Hence, there is a need to reform current technologies and invent newer ones so that end-to-end QoS in future wireless mobile systems can be addressed. In the following sections we shall look into certain aspects of RRM schemes that address the provisioning of QoS in future mobile wireless communications.

FUTURE MOBILE COMMUNICATION SYSTEMS

In the context of future networks it is expected that:

- The system is capable of supporting high data rates and users are able to connect to the best network as long as the terminal power is switched on and experience minimal access delay.
- Providing services comparable to wired networks for applications such as interactive multimedia, voice over internet protocol (VoIP), network games, video conferencing, and so forth.
- Providing multi-service ubiquitously and pervasively in diverse environments from indoor, outdoor (low and high velocity), and to global broadband access (satellite).
- Deployment over heterogeneous environments of various access networks.
- Users are always connected to the most efficient and available access networks catering to their specific QoS and mobility requirements.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- The system would be an integral part of Internet infrastructure (based on Internet protocol (IP) technology).

Of course the above key attributes need significant high spectral efficiency and can only be achieved via innovative techniques in air interface design and by implementing coverage enhancement techniques in which current systems might not be able to support fully. Future wireless systems will be expected to comprise multitudes of wireless air interfaces, such as WLAN, 3G cellular, beyond 3G or 4G cellular, peer-to-peer (P2P), multihop relays, and so forth, where different ranges of cell or coverage size are being supported within an integrated wireless access umbrella.

Within the context of cellular systems such a trend is already in placed where the former is working seamlessly with WLAN (IEEE 802.11, 1999) to provide ubiquitous and pervasive access to mobile users. As future networks will evolve into an ever integrating environment, the concept of "optimised connectivity," whereby maximising user experience while minimising radio operator resources, will be the main feature of such an architecture. To exploit the increased capabilities and potentialities of the systems, efficient resource management strategies in key aspects of the evolution need to be taken into consideration. In the next section we will discuss some of the areas in RRM that need to be addressed for providing users with varying QoS requirements.

RESOURCE MANAGEMENT IN FUTURE SYSTEMS

Given the proliferation of Internet and IP technology, which move in parallel with the advancement of mobile communication systems, eventually more and more IP will be adopted in beyond 3G wireless technologies. Future mobility service networks will be Internet-oriented and peer-to-peer architecture with no central mobile switching centre. Hence, we envision that almost all applications in future will be functionally dependent on IP. The main reason for integrating IP with a communication system is that both technologies can enhance heterogeneous support and interoperability of lower-layer technologies, that is, to continue building a network of networks. Since the QoS over IP-based applications is inherently unreliable, here we will look into the additional measures to support existing RRM schemes, which are crucial for providing QoS in such future wireless networks.

QoS Provisioning in IP Networks

To support the provisioning of QoS in an IP-based wireless communication system, the Internet Engineering Task Force (IETF, 2006) has proposed techniques such as reservation protocol (RSVP), integrated services protocol (IntServ) and differentiated services protocol (DiffServ) for QoS provisioning in IP networks. However, these models have been designed to work for wired networks in static environments and the aim now is to identify modifications to make them suitable for wireless networks. As mobility is becoming more and more popular on a daily basis, modifications to the IntServ and DiffServ models for use with wireless networks are discussed in length in Johnson et al. (2004), Kan et al. (2001), Lopez et al. (2001) and Moon et al. (2004). In addition the IETF Working Group is also aiming to provide seamless mobility across access routers and even domains, and the MIND project funded by the European Commission is also studying open IP-based mobile wireless network architecture to allow inter-operability between networks.

Traffic Control

Traffic control encompasses the application of scientific principles and technology in planning of network capacity under QoS guarantee and efficient transferring of information. The need to allocate and balance resources among different traffic classes to accomplish the best use of network resources is a crucial traffic engineering problem. The major objective of such a problem is to improve network performance while maintaining the QoS requirements through the optimisation of network resources. In order to support end-to-end QoS resources, such as bandwidth, scheduling time and buffer-space, more research needs to look into Internet traffic management schemes.

Presently traffic is usually routed on the shortest path through a network for both "best-effort service" and "guaranteed service." This is the case even if the shortest path is overloaded and there exist alternative paths that are underutilised. Hence, efficient optimisation techniques can be applied to IP networks in order to better utilise network resources and to avoid congestion by balancing load over several paths (Gunnar et al., 2005). In the context of treating the dynamics of Internet traffic management such as call admission control, self learning techniques (Gallador et al., 2001) could be used for identifying future traffic trends. Such techniques usually correlate current traffic variations of a given parameter to previous records of the parameter, and then predict favourable scenarios for transmission ahead of time.

Spectrum Management

In the future, we envisage spectrum will become an increasingly scarce resource and will force many operators to build unnecessary and expensive infrastructure with many base stations. Moreover, growing wireless users will only aggravate this problem, and hence spectrum shortage will have the potential to inhibit future growth. Provided governments allocate abundant spectrum to the wireless industry or to unlicensed bands for free use, growth will be slowed. Another way to reduce the problem is to develop new technology, enabling a more efficient use of the existing spectrum.

In the context of spectrum efficiency, several bandwidth partitioning strategies (Iraqi et al., 2000) have been designed to allocate bandwidth "fairly" for different traffic classes. Among them are complete sharing (CS) and complete partitioning (CP) (also known as mutually restricted access or MRA) strategies, and those in between are known as hybrid strategies. In CS schemes, all traffic classes share the entire bandwidth, with the Achilles heel being that a temporary overload of one traffic class would immediately offset and degrade the connection quality of all other traffic classes. As for CP techniques, the bandwidth is partitioned into different portions with each portion catering to a particular class of traffic. However such a technique is not economical feasible if the actual bandwidth demand for a particular traffic class is greater than the predicted bandwidth demand. As a compromise the hybrid CS-CP techniques have been proposed in which the bandwidth can be allocated dynamically to match the varying traffic classes. Nevertheless, finding the optimal partitioning point is not easy since the model is an NP-complete graph colouring problem, and hence only heuristic solutions on homogenous traffic have been proposed (Habib, 1997). In the future the bandwidth allocation problem is becoming even more complex, when wireless multimedia networks would then carry an integrated non-homogenous traffic in a mobile environment and, as such, new techniques have to be discovered to use the radio spectrum efficiently.

As radio spectrum is considered a public resource, an alternate method for government is to introduce spectrum trading (Hills et al., 2004) instead of licensing it to network operators for a certain amount of money. Since spectrum trading would inevitably involve monetary exchange, it can then increase market awareness of prevailing spectrum value. Hence market players can use this information to determine whether they are utilising the spectrum efficiently, whilst government appointed regulators may use prevailing market prices to assess whether spectrum allocations are indeed optimally being used. However, spectrum trading also introduces a host of uncertainties, the most critical of which is political. In the process of assigning spectrum rights, the government surrenders control of the spectrum to the market forces. This allows the possibility of a dominant player accumulating significant amounts of spectrum and therefore controlling the spectrum market.

Mobility Management

As more traffic in future wireless networks are expected to be mostly generated by multimedia applications, hence wireless networks need to provide adequate support for multimedia services and support user mobility with extended geographic coverage. Unfortunately multimedia applications often require a dynamic amount of bandwidth and, to guarantee QoS for such bandwidth-greedy applications when used over a wireless link, current schemes for supporting such services have to be reviewed and new resource management solutions have to be proposed.

A critical aspect of guaranteeing QoS support in providing seamless access under changing radio resource conditions is handoff or handover between the mobile terminal and the network. As a result handoff is related to access, radio resources and network resources, and has a direct impact on system capacity and performance. Presently most of the handoff techniques (Iraqi et al., 2000; Periyalwal et al., 2003) attempt to anticipate handoffs by using measurements of signal strengths and prior knowledge of mobility patterns and, in essence, have not entirely addressed the QoS requirement end-to-end. It is anticipated that with the profusion of future support services in tandem with the integration of multiple access networks, further standardisation towards an optimal mobility handoff solution in needed for all levels of QoS.

Another critical aspect of mobility management problem is location discovery (Iraqi et al., 2000; Periyalwal et al., 2003), which stems from navigation to context-aware applications in ubiquitous computing. As expected, the location requirements vary across applications, resulting in the development of a host of diverse sets of technologies and algorithms. In cellular telephony systems, for example, knowledge of handset locations is primarily needed for routing calls to other mobile users. Such a location is found either by measuring the received signal strength transmitted by the base stations or by measuring the round-trip signal propagation time between a handset and a set of base stations. Although cell-level localisation is sufficient for routing calls to mobile handset users, future wireless network applications require more precise handset locations to enable RRM schemes to function optimally. Recent research techniques in this direction are described in length in Savides et al. (2004).

Power Control and Antenna Beamforming

The physical layer of all wireless networks embodies a number of parameters that can be controlled for improved performance. Such parameters include modulation, transmit power, spreading code and antenna beams. By controlling these transceiver parameters adaptively and in an intelligent manner, one can increase the capacity of the system tremendously. In this section we only consider two synergistic parameters—transmit power and antenna beamforming—as they are intuitively the easiest to exploit and perhaps the most studied. The benefits of antenna beamforming include reduced interference due to narrower beamwidth, longer range due to higher signal to noise and interference ratio, and improved resistance to jamming. The benefits of power control on the other hand include reduced interference and lower energy consumption. In short, we envisage future techniques in interference mitigation in wireless networks would combine both power control and antenna beamforming to enable higher capacity due to increased spatial reuse, lower latency, better connectivity, longer battery lifetime, and better security (minimise eavesdropping). However, much work needs to done in analysing the theoretical network capacity of using both power control and beamforming (Gupta et al., 2000). In addition, a thorough study of other physical layer parameters such as modulation, spreading codes and others can be exploited to be used in conjunction with power control and beamforming, so as to open up new research frontiers.

Channel Allocation and Routing Mechanism

One of the most common resource management issues faced by service providers is channel allocation, which involves frequency planning, channel reuse schemes and network capacity optimisation. Channel allocation on its own influences the network performance as it limits interference while maximising the system capacity. On the other hand, a routing algorithm's main objective is to secure a communication link between the source and the destination nodes. In the future Wi-Fi or WiMAX mesh networks scenario, sources and destinations nodes may not be within direct transmission range of each other, and hence routing algorithms must be able to discover multihop routes between the sources and destinations so that communication between these nodes is possible. Although there are many routing and channel allocation algorithms (Chiasserini et al., 2000; Chuang, 1993) for a variety of topological scenarios, not much study has been done to exploit both channel allocation and routing algorithms as a single entity to serve the needs of future wireless community. Of course it is unlikely that a single algorithm exploiting both routing and channel allocation can solve the needs of every conceivable future wireless network scenarios. Hence, certain algorithms will likely to perform well in networks having a set of key attributes while others will perform better in networks having a different set of key attributes. Such a problem indeed highlights the importance of identifying sets of characteristics of a network in order to maximise the efficiency of the class of algorithms concerned.

CONCLUSION

At the dawn of the new millennium we have witnessed a fantastic growth of mobile communications particularly at the access network level. With the ever integration of Internet access and multimedia services, more and more sophisticated resource management mechanism are therefore needed to handle the large volume of heterogeneous traffic. In this article we have discussed some RRM issues that are paramount in providing ubiquitous and pervasive mobile broadband wireless access for the future wireless community. Of course there are many more challenging and open issues and we do not claim to have addressed them all. In addition, the popularity of mobile ad hoc wireless networks (Sim et al., 2006) is rapidly increasing and is soon to be a permanent fixture in wireless communication systems. As such it is envisioned that RRM schemes will receive more attention since the QoS provisioning is more challenging in ad hoc wireless networks due to their decentralised nature and lack of infrastructure. Hence, there will be further challenges of coping with uncertainty in wireless networks, such as time varying wireless links, user mobility, routing, traffic loads, security and so forth. It is hoped that future multimedia networking will reach to a point of having the capability of self-organising appropriate RRM schemes optimally to adapt to the required network environment and also to the type of multimedia traffic.

REFERENCES

Chiasserini C. F., & Rao, R. R. (2000). Routing protocols to maximize battery efficiency. In *Proceedings of IEEE MILCOM*. Los Angeles, CA.

Chuang L. C. I. (1993). Performance issues and algorithms for dynamic channel assignment. *IEEE Journal on Selected Areas in Communications*, *11*(6), 955-963.

Gallador J. R., Makrakis D., & Angulo M. (2001). Dynamic resource management considering real behaviour of aggregate traffic. *IEEE Transactions on Multimedia*, *3*(2), 177-185.

Gunnar A., Abrahamsson H., & Söderqvist M. (2005, In Press). Performance of traffic engineering in operational IP networks: An experimental study. IPOM 2005.

Gupta, P., & Kumar, P. R. (2000). The capacity of wireless networks. In *IEEE Transactions on Information Theory* (IT-46) (pp. 388-404).

Habib, I. (1997). Bandwidth allocation in ATM networks. *IEEE Communications Magazine*, *35*, 120-121.

Hills, T., & Pow, R. (2004). *Spectrum trading and liberalisation: New threats and opportunities for telecoms business models*. Analysys Consulting Report.

IEEE 802.11. (1999). *Wireless LAN medium access control (MAC) and physical layer (PHY) specifications*. Retrieved from http://standards.ieee.org/getieee802/802.11

IEEE 802.20. (2006). *Mobile broadband access*. Retrieved from http://grouper.ieee.org/groups/802/20

IETF. (2006). *The Internet Engineering Task Force*. Retrieved from http://www.ietf.org/

Iraqi, Y., & Boutaba, R. (2000). Resource management issues in future wireless multimedia networks. *Special issue on the Management of Multimedia Networking*, *9*(3-4), 231-260.

Johnson, D., & Perkins, C. (2004). *Mobility support in IPv6*. Internet Engineering Task Force. Retrieved from http://rfc3775.x42.com/

Kan, Z., Zhang, D., Zhang, R., & Ma, J. (2001). QoS in mobile IPv6. In *Proceedings of International Conferences on Info-tech and Info-net 2001* (Vol. 2) (pp. 492-497).

Lopez, A., Manner, J., Mihailovic, A., Velayos, H., Hepworth, E., & Khouaja, Y. (2001). A study on QoS provision for IP-based radio access networks. In S. Palazzo (Ed.), *Evolutionary trends of the Internet: Thyrrhenian International Workshop on Digital Communications (IWDC 2001)* (LNCS). Springer-Verlag GmbH.

Moon, B., & Aghvami, A. H. (2004). Quality of service in mobile IP networks. In *The mobile Internet: Enabling technologies and services*. Apostolis K. Salkintzis (Editor). CRC Press.

Periyalwal, S., Hashem, B., Senarath, G., Au, K., & Matyas, R. (2003). Future mobile broadband wireless networks: a radio resource management perspective. *Wireless Communications and Mobile Computing*, *3*, 1-13.

Savvides, A., & Srivastava, M. B. (2004). Location discovery. In S. Basagni, M. Conti, S. Giordano, & I. Stojmenovic (Ed.), *Mobile Ad Hoc Networking*. IEEE Press.

Sim, M.L., Chin, C.M., & Tan, C.M. (2006). Mobile ad-hoc networks. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce*. Hershey, PA: Idea Group Publishing.

Thelander, M. W. (2005). *The 3G evolution*. CDG White Paper.

UMA. (2006). *Unlicensed mobile access*. Retrieved from http://www.umatechnology.org/

Wi-Fi Alliance. (2006). Retrieved from http://www.wi-fi.org/

WiMAX Forum. (2006). Retrieved from http://www.wimaxforum.org/

KEY TERMS

3G: Third generation mobile network.

4G: Fourth generation mobile network.

MIND: Mobile IP-based network development, IST-2000-28584.

Mobile-Fi: Mobile broadband wireless access.

NP: Non-deterministic polynomial time.

WLAN: Wireless local area network.

Wireless Fidelity (Wi-Fi): It is also referred as IEEE802.11. It is a set of standards that set forth the specifications for transmitting data over a wireless network.

Worldwide Interoperability for Microwave Access (WiMAX): It is frequently referred to as the IEEE 802.16 wireless broadband standard. It is initially designed to extend local Wi-Fi networks across greater distances such as a campus, as well as to provide last mile connectivity.

P2P Models and Complexity in MANETs

Boon-Chong Seet
Nanyang Technological University, Singapore

Chiew-Tong Lau
Nanyang Technological University, Singapore

Wen-Jing Hsu
Nanyang Technological University, Singapore

INTRODUCTION

Computing systems are playing an essential role as an indispensable nervous system of modern society. In recent years, peer-to-peer (P2P) computing has gained significant attention from both industry and research communities (Barkai, 2001). A key attraction of P2P systems is their ability to scale without requiring expensive and powerful servers, primarily because P2P systems work by distributing the functionality and harnessing the resources across a large number of independent peers. In addition to having high scalability, such systems are also inherently robust and fault tolerant since there is no centralized server, and the network is inherently self-organized. Today, P2P technology has been widely embraced by Internet users and has seen highly successful applications in such areas as digital data sharing, voice over peer-to-peer, distributed computing, and distributed storage (Miller, 2001). With the advent of wireless technology, the number of mobile users has increased tremendously over the years. It is envisioned that a significant portion of future users of P2P systems would be based on mobile ad-hoc networks (MANETs), which are dynamic networks formed by peers with no support of fixed infrastructure networks (Ramanathan & Redi, 2002; Perkins, 2001). As contemporary P2P models are designed mainly for the Internet, research is needed to examine their viability for P2P computing in a mobile environment.

This article first overviews some of the most common P2P architectures in use today by Internet users, and then proceeds to evaluate and discuss each of their architectural strengths and weakness for MANET through a qualitative complexity analysis.

INTERNET P2P MODELS

Internet P2P systems are a popular paradigm for data exchange in a decentralized manner between computer users across the Internet. A P2P network over the Internet is a highly dynamic overlay network where users join and leave the network frequently. Numerous P2P systems that have been proposed for the Internet in recent years may be broadly classified into two categories: *unstructured* and *structured* (Lua, Crowcroft, Pias, Sharma, & Lim, 2005).

Unstructured P2P systems do not have tight control of the overlay topology. The network is typically formed by nodes in a random manner, and thus the object or data that one wishes to locate could be anywhere in the system. A query has to be flooded through the network to search for peers that store the desired object. Besides a high amount of signaling traffic, the search may be prematurely terminated if the object could not be found within the query lifetime, often specified in terms of number of hops.

Structured P2P systems, on the other hand, impose a certain structure on the overlay topology and control of data placement to enable more efficient and reliable location of objects in a bounded number of hops through the use of distributed hash table or DHT (Stoica et al., 2003). In DHT, every object is mapped to some peer. A description (key) containing a link (value) to where the object could be found is then stored at the peer to which the object is mapped. This forms a deterministic relationship between objects and peers, and consequently, the query need not be flooded but routed directly to the peer responsible for storing the object location.

Until recently, Internet P2P systems (including both unstructured and structured systems) assumed all peers are equal and uniform in resource capacity. System functionality is thus distributed without considering real-world heterogeneity of peer capabilities. For example, some peers have less memory and slower processor than others, but they perform the same role and responsibility as other peers with higher capabilities. This results in instances of bottlenecks in performance due to very limited resources of these peers. To account for and even exploit the heterogeneity of peer capabilities, the notion of *super peers,* or nodes more well provisioned in terms of resources, have recently been introduced and advocated (Yang & Garcia-Molina, 2003; Singh, Ramabhadran, Baboescu, & Snoeren, 2003). Super peers take on a greater role and responsibility by serving as local

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

search hubs that manage (receive and resolve) object queries of ordinary peers. Each super peer in turn communicates with other super peers as equals in a pure P2P way. As a result of clustering heterogeneous devices and elevating certain well-provisioned nodes to the role of super peers, the impact of performance bottlenecks presented by some less capable peers could be minimized.

P2P IN MANETs: A COMPLEXITY ANALYSIS

P2P systems for the Internet are designed with a decentralized architecture, which makes them potentially suited to the infrastructure-less MANET environment. However, as each overlay link could consist of multiple hops in the underlying physical network, a dynamic MANET topology due to peer mobility could pose significant problems for data management, particularly in structured P2P systems (Hsiao & King, 2005). The lower capacity of MANET compared to the Internet and the resource constraints of mobile devices further limit the usability of P2P architectures that have high traffic overhead.

This section provides a complexity analysis of the asymptotic cost and performance of three popular Internet P2P architectures—unstructured super peer, structured non-super peer, and structured super peer—in a mobile environment. The cost measure is the message complexity, or number of hop-wise message transmissions required to perform a specific operation such as node join, node leave, and object query. The performance measure is the time complexity or number of time steps needed to perform an operation.

Unstructured Super Peer

The popular file-sharing application Kazaa (Liang, Kumar, & Ross, 2005) and voice-over-IP (VoIP) client Skype (Baset, & Schulzrinne, 2006) are prominent examples of commercial-grade P2P systems based on unstructured super peer architecture. As these systems are originally designed for wired networks, as a first step, we need to imagine how such a system may be used in a mobile environment. For super peer-based architectures, we reason that the most natural choice of an underlying physical network would be a cluster-based MANET (Yu, & Chong, 2005), since the cluster-heads are ready candidates for super peers in the P2P system.

We assume each ordinary peer (mobile user) would know a super peer (cluster-head) through the underlying clustering mechanism. When a mobile user *u* wishes to join, it sends a *join* message to its super peer *P* with a list of objects it wishes to share, and records *P* as the super peer through which it joins the system. Upon receiving, *P* adds the object list of *u* into its index. When *u* wishes to leave, it sends a *leave* message to its super peer. If the current super peer is not *P* due to node mobility—that is, *u* is now associated with a different super peer in another cluster—the super peer forwards the message to *P*. Upon receiving, *P* deletes the object list of *u* from its index. When *u* wishes to query for an object, it sends a *query* message to its super peer. If the super peer knows the object, it replies to *u* immediately with the object location. Otherwise, it broadcasts the query to other super peers (via cluster gateways). Assume from the reply, *u* knows that a mobile user *v* is holding the object. *u* then proceeds to contact and retrieve the object from *v*.

Having outlined the join, leave, and query operation of the unstructured super peer system in a MANET environment, we may now begin our analysis. Consider a network of *n* nodes, of which a fraction *m* (where $0 < m < 1$) are super peers. Thus, we have mn super peers and $(1-m)n$ ordinary peers. Let us denote n_s and n_o as the number of super peers and ordinary peers, respectively. If *m* is a constant, then n_s and n_o would increase in proportion to *n*. Under constant node density, *n* itself would increase in proportion to *A*, the network area—that is, $n \propto A$.

If the communication pattern is generally uniform, the average path length *L* or number of hops between any pair of nodes is expected to grow with the spatial diameter of the network, which in turn grows with the square root of the area, or equivalently $\sqrt{n}$ (Li, Blake, Couto, Lee, & Morris, 2001). Similarly, we expect path length L_s between any pair of super peers, and path length L_o between any pair of ordinary peers, to grow with $\sqrt{n}$.

During a join, the *join* message from a mobile user to its super peer travels a path length of *c*, where *c* is a positive constant denoting the number of hops to a cluster head, which depends only on the clustering scheme used—that is, 1- or multi-hop clustering (Yu & Chong, 2005). The join process thus has a constant message and time complexity of $O(c)$. During a leave, the *leave* message travels under worst case a path length of $c+L_s$: mobile user → its current super peer → the super peer that originally hosts its object list, leading to a message and time complexity of $o(\sqrt{n})$ for the leave process.

The analysis for the query process is slightly more involved, thus we present for only the most common case of single-hop clustering: $c = 1$. First, consider a mobile user sending a *query* message to its super peer. If the super peer does not know the object, it broadcasts the query to other super peers. We denote a fraction *q* of the ordinary peers as gateways that rebroadcast the query to other super peers. If *k* is the fraction of super peers that reply with the object location, there would be kn_s replies and $(1-k)n_s$ rebroadcast queries by the super peers. Each reply would travel a path length of L_s+1: replying super peer → super peer of mobile user → mobile user. Summing up, this gives a total message cost of:

$$1+q(1-m)n+(1-k)n_s+kn_s(\sqrt{n}+1) \qquad (1)$$

Since n_s is a constant fraction of n, the message complexity of the query process would scale as $O(n^{3/2})$, while the corresponding time complexity would be in the order of $o(\sqrt{n})$.

Structured Non-Super Peer

Distributed lookup service Chord (Stoica et al., 2003) is a prominent example of a system with structured non-super peer architecture. For non-super peer-based architectures, which deem all peers as being equals, the most natural choice of an underlying physical network would be a non-hierarchical MANET. In the following, we outline the steps involved in the join, leave, and query processes of Chord.

We assume each peer knows its 1-hop neighbors, for example via a hello mechanism. When a mobile user *u* wishes to join, it sends a *join* message to a neighbor *v* that has already joined the system. Upon receiving, *v* helps *u* to find its immediate successor, which may take up to log *n* steps to complete, where *n* is the number of nodes in the network. Upon knowing its successor *s*, *u* informs *s* that it has a new predecessor, and *s* in turn transfers a portion of its keys (object tuples) to *u*. On the other hand, *s*'s original predecessor *p* becomes aware of *u* through a stabilize process, and proceeds to inform *u* that it is *u*'s predecessor. Having set up its successor and predecessor pointers correctly, *u* needs to build its finger (routing) table by sending up to *m-1* discovery messages, where *m* is the number of table entries and *s* being in the first entry. *u* then begins to key-insert (publish) for each object it wishes to share, with each insertion taking up to log *n* steps to complete.

When *u* wishes to leave, it sends a *leave* message to its successor and predecessor. In the message to its successor, it also includes the keys it wishes to hand over. Before leaving, *u* performs a key-delete (withdraw) for each object it has shared, with each deletion taking up to log *n* steps to complete. When *u* wishes to query for an object, it sends a *query* message to a node in its finger table whose *id* immediately precedes that of the object. This node in turn forwards to one of its own finger nodes in the same way, and repeats until the key is found, which again may take up to log *n* steps. Finally, assuming from the key that *u* knows a mobile user *v* is holding the object, *u* then proceeds to contact and retrieve the object from *v*.

In this analysis, we similarly consider a network of *n* nodes, with path length *L* between any pair of nodes growing with $\sqrt{n}$, under the assumption of constant node density. During a join, the *join* message from a mobile user *u* may be forwarded log *n* times, followed by three message exchanges to configure the successor and predecessor pointers of *u*. Next, *u* sends *(m-1)* messages to build its finger table, and up to *f* log *n* messages to key-insert its objects, where *f* is the number of objects. If each of these messages travels a path length of *L*, this gives a total message cost of:

$$1+[\log n+3+(m-1)+f\log n]\sqrt{n} \qquad (2)$$

The message complexity of join process is thus $o(\sqrt{n}\log n)$. To compute the time complexity, we assume the transmission of each key-insert message to be separated by only a small time constant ε. The total time incurred is therefore: $1+[\log n+3(m-1)]\sqrt{n}+\varepsilon(f)+\sqrt{n}\log n$, which still has a complexity of $o(\sqrt{n}\log n)$.

During a leave, the total amount of messages and time incurred by *u* is $(2+f\log n)\sqrt{n}$ and $2\sqrt{n}+\varepsilon(f)+\sqrt{n}\log n$, respectively. This results in a message and time complexity of $o(\sqrt{n}\log n)$ for the leave process. For the query process, it could be easily seen that its message and time complexity is also of order $o(\sqrt{n}\log n)$.

Structured Super Peer

This section evaluates a system that combines the super peer concept with structured object location (Mizrak, Cheng, Kumar, & Savage, 2003). In this architecture, the super peers manage data in a structured manner, while ordinary peers query and receive replies from their super peers, similar to conventional super peer systems. As in our analysis for unstructured super peer systems, we assume the underlying physical network would be a cluster-based MANET, where cluster-heads perform the role of super peers.

Similarly, we consider a network of *n* nodes, of which n_s are super peers and n_o are ordinary peers. During a join, the *join* message from a mobile user to its super peer travels a path length of *c*. If the message contains a list of *k* objects to be shared, the super peer may key-insert for up to *k* times, incurring a total cost of $c+k\sqrt{n}$ messages: mobile user → its current super peer → the super peer(s) responsible for storing the key(s). Summing up, it could be seen that message complexity of the join process is $o(\sqrt{n})$. In terms of time complexity, we have a total time incurred of $c+\varepsilon(k)+\sqrt{n}$. The time complexity is thus also of order $o(\sqrt{n})$.

Since the leave process is similar to the join process—that is, the mobile user sends a *leave* message to its super peer, which may key-delete for up to *k* times—the leave process is found to have the same message and time complexity of $o(\sqrt{n})$. During a query, both the *query* message and *reply* message may travel up to a path length of $c+\sqrt{n}$: mobile user ↔ its current super peer ↔ the super peer that holds the key. Thus, the query process has a message and time complexity of $o(\sqrt{n})$.

Table 1. Summary of message and time complexity

P2P Models	Message and Time Complexity		
	Join	Leave	Query
Unstructured Super Peer	$O(c)$	$O(\sqrt{n})$	$O(n^{3/2})$ (message) $O(\sqrt{n})$ (time)
Structured Non-Super Peer	$O(\sqrt{n}\log n)$	$O(\sqrt{n}\log n)$	$O(\sqrt{n}\log n)$
Structured Super Peer	$O(\sqrt{n})$	$O(\sqrt{n})$	$O(\sqrt{n})$

CONCLUSION

Table 1 summarizes the complexity of join, leave, and query processes of three P2P models. Generally it is observed that a super peer model is efficient in its *join* and *leave* process. Thus, it may perform well in a churn-intensive environment—that is, frequent join and leave of mobile nodes. However, without a structured method for object location, its query process is potentially costly since the query message could be broadcast to all super peers. Therefore, it is a challenge for unstructured super peer systems to perform under a query-intensive workload.

On the other hand, with structured object location, the *query* process could be more efficient, since the query message would be ‘directed’ and not ‘flooded’. This is achieved by maintaining some state on each node to serve as ‘direction’ pointers. However, in a high-churn environment, the traffic to maintain state could be high: some state needs to be inserted or deleted whenever a node joins or leaves. Without super peers, this traffic could be even higher, as it takes more steps to find the correct node to insert or delete the state (typically up to *log n* steps). Each step traverses an overlay link that potentially spans the diameter of the underlying MANET, leading to a very high forwarding traffic. A recent work (Crammer, & Fuhrmann, 2005) on overlay topology construction based on proximity of nodes in the underlying physical network could potentially address this problem.

The structured super peer model appears to be relatively well performed in all three processes. This could be due to its combining of strengths from both super peer design (efficient join/leave) and structured object location (efficient query), and is thus a promising P2P model to be further explored, particularly for a cluster-based MANET.

REFERENCES

Barkai, D. (2001). *Peer-to-peer computing: Technologies for sharing and collaborating on the Net.* Intel Press.

Baset, S.A., & Schulzrinne, H. (2006). An analysis of the Skype peer-to-peer Internet telephony protocol. *Proceedings of IEEE INFOCOM 2006,* Barcelona, Spain.

Cramer, C., & Fuhrmann, T. (2005). Proximity neighbor selection for a DHT in wireless multi-hop networks. *Proceedings of the 5th IEEE International Conference on Peer-to-Peer Computing,* Konstanz, Germany.

Hsiao, H.-C., & King, C.-T. (2005). Mobility churn in DHTs. *Proceedings of the 1st International Workshop on Mobility in Peer-to-Peer Systems,* Columbus, OH.

Li, J., Blake, C., Couto, D., Lee, H. I., & Morris, R. (2001). Capacity of ad hoc wireless networks. *Proceedings of the 7th ACM International Conference on Mobile Computing and Networking,* Rome, Italy.

Liang, J., Kumar, R., & Ross, K. W. (2005), The Kazaa overlay: A measurement study. *Journal of Computer Networks (Special Issue on Overlays), 49*(6).

Lua, E. K., Crowcroft, J., Pias, M., Sharma, R., & Lim, S. (2005). A survey and comparison of peer-to-peer overlay network schemes. *IEEE Communications Surveys and Tutorials, 7*(2), 72-93.

Miller, M. (2001). *Discovering P2P.* New York: John Wiley & Sons.

Mizrak, A. T., Cheng, Y., Kumar, V., & Savage, S. (2003). Structured superpeers: Leveraging heterogeneity to provide constant-time lookup. *Proceedings of the IEEE Workshop on Internet Applications,* San Jose, CA.

Perkins, C.E. (2001). *Ad hoc networking.* Boston: Addison-Wesley.

Ramanathan, R., & Redi, J. (2002). A brief overview of ad hoc networks: Challenges and directions. *IEEE Communications, 40*(5), 20-22.

Singh, S., Ramabhadran, S., Baboescu, F., & Snoeren, A. (2003). The case for service provider deployment of super-peers in peer-to-peer networks. *Proceedings of the International Workshop on Economics of Peer-to-Peer Systems,* Berkeley, CA.

Stoica, I., Morris, R., Liben-Nowell, D., Karger, D. R., Kaashoek, M. F., Dabek, F., & Balakrishnan, H. (2003). Chord: A scalable peer-to-peer lookup protocol for Internet applications. *IEEE/ACM Transactions on Networking, 11*(1), 17-32.

Yang, B., & Garcia-Molina, H. (2003). Designing a super-peer network. *Proceedings of the 19th International Conference on Data Engineering,* Los Alamitos, CA.

Yu, J. Y., & Chong, P. H. J. (2005). A survey of clustering schemes for mobile ad hoc networks. *IEEE Communications Survey and Tutorials, 7*(1), 32-48.

KEY TERMS

Asymptotic Cost: The algorithmic cost as the problem size grows very large.

Cluster Gateway: A node that enables communication between two or more cluster heads, either by being a member of more than one cluster, or by being adjacent to a gateway of another cluster.

Complexity Analysis: The analysis of how an algorithm scales, for example, in terms of number of messages and time required, with increasing problem size.

Fixed Infrastructure: A fixed topology of network devices such as routers and base stations that does not change during the lifetime of a connection.

Overlay Network: A logical network of nodes layered on top of a physical network in order to perform application-level services.

Physical Network: A network comprised of nodes and the actual communication links between them.

Query Lifetime: A threshold count often specified in terms of number of hops a query message can be forwarded, beyond which the message is dropped.

Partial Global Indexing for Location-Dependent Query Processing

James W. Jayaputera
Monash University, Australia

INTRODUCTION

Mobile computing devices, which are small-size computers, enable mobile users to run their applications to access information via wireless communication anywhere at anytime (Barbara, 1999; Dunham & Kumar, 1998). Applications such as stock or banking activities, weather forecasting, entertainment, and navigation systems have become increasingly popular in recent years (Mobile Computing Horizontal Applications Index, 2006). This implies that people can still access information through their mobile devices without worry about their current location. This situation enables the capability of mobile users to utilize information services easily. However, these mobile computing devices have limited battery power and bandwidth. Hence, speeding up the processing of information access and retrieval is a major challenge to minimize the utilization of battery power and available bandwidth.

In performing the many popular applications mentioned earlier, location-dependent information services is one type of mobile computing capability for providing information based on the current location of mobile users (Baihua, Lee, Xu, & Lee, 2002; Dunham & Kumar, 1998; Waluyo, Srinivasan, & Taniar, 2005). This means that the relaying back of a query result changes according to the current location of the mobile user. If, for example, a mobile user would like to locate the closest restaurant, a list of the closest restaurants depends on the current location of the mobile user. The records in the list change as the location of the mobile user changes.

The objects residing in several servers are partitioned by their locations. This indicates that every object is unique to every server. The objects can be classified into *static* and *dynamic.* An example of a static object is an ATM (automatic teller machine) or building. In contrast, an example of a dynamic object is a moving car or running thief. In this article, we concentrate on static objects.

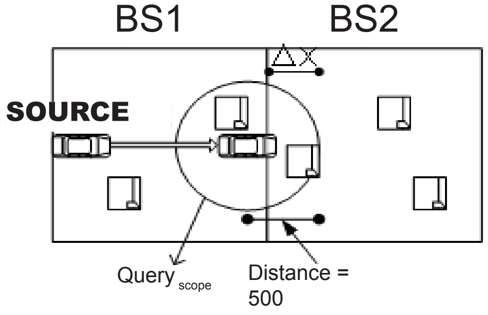
To speed up the searching process, these objects are indexed and the indexes are put into a data structure. There are several popular data structures, such as B+-Tree (Elmasri & Navathe, 2004), Quadtree (Samet, 1984; Tayeb, Ulusoy, & Wolfson, 1998), and R-Tree (Guttman, 1984; Manolopoulos, Nanopoulos, Papadopoulos, & Theodoridis, 2005).

In a mobile computing environment, a specific area is called a *cell* and is served by a base station (BS). In a simple situation, a requested query scope can be differentiated into two types—that is, *within* or *crossing* the current cell. In the former, the access time to process a query within a current cell depends on the queuing and processing times. In the latter, the access time to process a query crossing the current cell is similar to the first case, but the queuing time of each cell varies. Here, the terms *queuing* and *crossing* are used interchangeably.

Figure 1 shows an overview of LDIS in a simple situation. For example, a user sends a *location-dependent query* (LDQ) while driving—that is, it asks for data which can be found within the current location of the user. The requested data is called *location-dependent data* (LDD) (Jayaputera & Taniar, 2005). The LDQ sent is to find some vending machines within 500 meters from the current location.

The query is received by a base station, which forwards the message between the wireless and wired network for a specific area, is called a *cell* (Goodman, 1998). We assume that a BS is connected to a single server. From the figure, the query scope is represented by a circle which shows the search area of the query. Before the server starts the searching process, it needs to know the location of the mobile user from the GPS (global positioning satellite) (Getting, 1993). We assume that the location information taken from the satellite is accurate. Once the recipient's location is received, the searching process begins. In our example, the query scope crosses the area of BS1, which needs to ask for the information from BS2 on behalf of the mobile user. Once BS1 has the requested information, it forwards the information to the mobile user. The correct answer to this query should be the information about vending machines located inside the circle.

Figure 1. Overview of LDIS



Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

In a complex situation, the scope of a requested query might cross a number of cells. The main consideration is that the scope of a requested query in a complex situation clearly differentiates between two types of scopes mentioned earlier: within or crossing the current cell.

This article concentrates on how to reduce the access time to retrieve information about static objects in a complex situation. The aim of this article is not to discuss concurrent search trees with their related problems and solutions, but to investigate how the practice of data/table partitioning in parallel databases can be applied in a mobile computing environment. As an example, the first query would be "Tell me a restaurant within 500 meters" and the second query: "Tell me a restaurant within 400 meters." Two mobile devices, which are located nearby, receive the results of these two queries, and the queries' scopes cross the current cell. In this scenario, the BS of the current cell needs to forward the request to the crossing cells. We assume that each BS has knowledge of its own service area and its neighboring cells. Requesting the same data from neighboring cells is not efficient since the process depends on some factors, such as the processing and waiting times on each neighboring cell. In the LDIS environment, a short access time is an important issue since the mobile user moves to a new location very frequently before receiving an answer from a server. We argue that accessing data locally needs to be improved and requesting data from other neighboring cells should be minimized.

The idea of this article is based on the parallel indexing concept (Taniar & Rahayu, 2002) in which an indexed object residing in a BS is either fully, partially, or not replicated to others BSs. Therefore, every server contains either partial or all indexes of other servers. In our proposed approach, whenever the requested results return from neighboring cells, we append the resulting items to the current cell. This implies that when the next user sends a request, the current cell needs to look up its own index first to verify if the data is in its local storage. If the data is not present, the current server sends a request to the neighboring cells on behalf of the client; otherwise, the current server directly sends the requested query to the client. We have evaluated our proposed approach and showed that the access time can be reduced by a factor of two.

The next section of this article describes some related work. We then describe our proposed work and the simulation model, and we compare the performance of our proposed technique to other techniques. Finally, we conclude the article and suggest future work.

RELATED WORK

Jayaputera and Taniar (2005) have shown the efficiency of using a square as a valid scope, which is an area in which the objects are valid for a certain time. They have shown that a square is the easiest shape to use.

Taniar and Rahayu (2002) discussed global indexing for a parallel database. The purpose of their approach was to have a global index of database servers. Whenever there is an update for an item in one server, the global index needs to be updated in all servers. The updating process for every server will increase the CPU load, where it can affect the performance of processing the LDQ. Therefore, this is not a practical choice to be applied directly for LDIS application.

B+-Tree (Elmasri & Navathe, 2004) has been widely known as one of the data structures for index which contains sub-tree and leaf nodes. A sub-tree is formed by a collection of non-leaf nodes. A non-leaf node contains up to *m* keys and *m+1* pointers to the nodes on the next level of the tree hierarchy. All nodes on the left-hand side of the parent node have key values less than or equal to the key value of the parent node. In contrast, the key values of the right-hand side nodes of the parent node are greater than the key values of the parent node. The most bottom nodes are called *leaf* nodes. Each leaf node contains up to *m* keys where every key has two pointers: to the actual data items and to the right-side neighboring leaf node.

Some researchers (Beckmann, Knegel, Schneider, & Seeger, 1990; Guttman, 1984; Sellis, Roussopoulos, & Faloutsos, 1987; Theodoridis & Sellis, 1994) used R-Tree and its variants to provide an efficient and dynamic index structure for spatial data. Some researchers (Sellis et al., 1987) have applied R-Tree to LDQ processing. R-Tree uses the *minimum bounding rectangle* (MBR) to group the closest objects together into a rectangle where every area has the least enlargement area. This data structure does not have a problem when a query range can be fitted into one rectangle. On the other hand, we need to have multi-way searching if a query range involves some rectangles.

Furthest away replacement (FAR) is one of the cache replacements proposed in Dunham and Kumar (2000). In their approach, they eliminated the data items located furthest away from the user and which will be evicted first since the users no longer need those objects. The other replacement method is using *timestamp,* which is the time at which data items are received by the cell. *Least recently used (LRU)* (Jelenkovic & Radovanovic, 2003) is used to eliminate data items that have the oldest timestamp.

PARTIAL GLOBAL INDEXING FOR OBJECT RETRIEVAL: PROPOSED ALGORITHM

In general, the data items are located in several cells; we assume that every cell has a single server, and users often

Figure 2. The proposed algorithm

```
Algorithm 1. FindNode (Tree, Key, Operation)

  N ← a root node of Tree
  if key is in the range of node N then
      if N is leaf Node then
          Execute operation insert/delete onlocal node
          if Node is overflow or underflow then
              Perform split/merge on leaf node
          end if
      else
          Locate child tree
          Perform FindNode(child, key, operation)
          if Node is overflow or underflow then
              Perform split/merge on leaf node
          end if
      end if
  else
      Locate child tree
      ItemsFromNeighbour ← GetItemFromNeighbour(cellID, key)
      Insert ItemsFromNeighbour to Local Node
      if Node is overflow or underflow then
          Perform split/merge on leaf node
      end if
  end if
```

request data items that are located in a single cell or multiple cells. Retrieving data items from multiple cells shows its limitations when a large number of queries are in queue. However, it is not appropriate to wait to retrieve all data items as the users will be moving to another location.

The proposed strategy is used to store the requested index of data items from neighboring cells during its first-time request. This process keeps going on until the next user requests the same data item that was requested before. When the B+-Tree reaches its maximum, some data items have to be invalidated. We invalidate the index of data items by using three cache replacements: *furthest away replacement, least frequently accessed,* and *least recently used.*

To simplify our discussion, we explain our proposed approach in conjunction with the B+-Tree. However, it is not limited with the B+-Tree data structure. It also can be used with variants of the R-Tree data structure. We adopt the insertion, deletion, and modification of the traditional B+-Tree. For further information, interested readers may refer to Elmasri and Navathe (2004).

Our approach is based on the global indexing concept. The original global index is designed based on the B+-Tree data structure in the context of parallel database systems (Taniar & Rahayu, 2002). We store indexes of the requested objects from neighboring cells locally since the size of an index is relatively small. In other words, we partially replicate the index of objects from neighboring cells, but we do not replicate the sub-tree of the remote tree if the index is already replicated. The idea is to store a complete list of indexes of all requested objects, which are located in several servers from neighboring cells. A partially replicated sub-tree can consume more processing time, which slows down the query processing.

Figure 3. Examples of object information retrieval from multiple cells

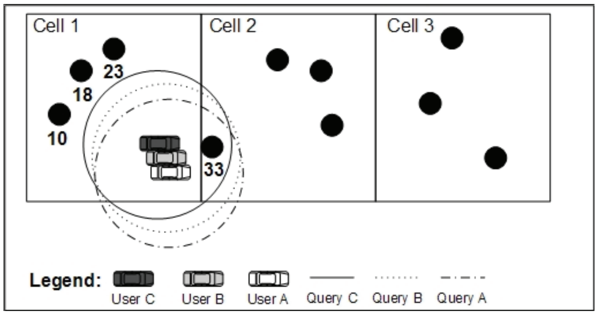


Figure 2 shows our proposed algorithm. In the beginning, we traverse from the root to the leaf nodes to find whether the index of the requested object is located in the B+-Tree of the current cell. If the index is found, the server forwards the information about the requested object to the user. Otherwise, the current cell forwards the request on behalf of the user to the neighboring cells. These servers search for the requested object in their local indexes. The indexes and information about the requested objects are then returned to the requesting cell.

Once the information about the requested objects is received from neighboring cells, the current cell performs two tasks: storing the indexes of the requested objects and sending the answer to the mobile user. The latter is straightforward, because the current cell sends the information to the mobile user directly. If the mobile user gets disconnected, then the current cell waits for a short amount of time before it resends the information about the object to the users. Otherwise, users need to resend the query again since their current locations might be different from the previous one.

The former task of the current cell is quite complex compared to the latter, since the server needs to insert the new indexes to the B$^+$-Tree index and restructure the tree. The tree is growing when there are new indexes inserted into the tree. Therefore, the physical memory space in a server (cell) will become full, while the process of storing new data items keeps going. It causes the server to run out of space. Therefore, we need to invalidate appropriate data items. In the first place, all objects that have the furthest distance from a cell boundary, rather than from the user, will be evicted. If there is more than one object that has the same distance, the data items that are least frequently accessed will be invalidated. If there are still remaining same numbers of frequently accessed data items, those with the older timestamp will be evicted.

Figure 3 shows a situation where there are three mobile users—A, B, C—that ask the same query, and their locations are close to each other and located in cell 1. The order of arrival time is A, B, and C respectively. When server 1 processes a query A (dot line), the server looks up the tree index in cell 1. Server 1 cannot find any requested objects in its tree index and so it checks whether the query scope is crossing the boundary. If it is, server 1 forwards the query to the neighboring cell, server 2. The queried object is located in server 2 and returned to server 1.

Server 1 forwards the results to the user and updates its tree index. For simplicity, the mobile user is connected while receiving the result and there is no failure in updating its tree index. Once processing of query *A* is completed, server 1 continues processing the next query, query *B* (dot). Server 1 only searches its tree index, because the object index is available locally as it was requested before. Thus, server 1 does not need to forward again the whole request from the user to its neighboring cell(s) in order to search the object. Server 1 only needs to request information about the requested object from the neighboring cells whenever they are not available in its tree. Then, server 1 processes the next query, query *C* (solid line). The processing steps for query *C* are similar to those for query *B*.

FUTURE TRENDS

The global indexing approach is often used in distributed parallel databases. However, this approach cannot be applied directly to answer a location-dependent query since a short time is needed to answer this query. Therefore, the partial global indexing approach, as a modification of the global indexing approach, is made to be used in retrieving objects for location-dependent queries. Further investigation will be done on how this algorithm can improve processing time to answer the location-dependent query.

CONCLUSION

This article proposes an approach called *partial global indexing.* Our proposed approach is similar to data/table partitioning in parallel databases. The difference is that we do not replicate database records from one cell to all other cells. The aim is to speed up the retrieval of object information from multiple cells by reducing the waiting time. In this article, we did not discuss concurrent search trees and their related problems.

In our approach, we used indexing to store the requested information from other cells to the cell where the mobile client currently resides. This implies that if there is no information entry in the local index regarding the object requested by a new mobile client, the server needs to retrieve the information from neighboring servers. The next time the same information is requested, the next mobile client only needs to find the items inside the cell where he currently resides.

REFERENCES

Baihua, Z., Lee, D. L., Xu, J., & Lee, W. C. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing, 1*(3), 65-72.

Barbara, D. (1999). Mobile computing and databases—A survey. *IEEE Transactions on Knowledge and Data Engineering, 11,* 108-117.

Beckmann, N., Knegel, H. P., Schneider, R., & Seeger, B. (1990, May). The R*-Tree: An efficient and robust access method for points and rectangles. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 19*(2), 322-331.

Dunham, M. H., & Kumar, V. (2000). Using semantic caching to manage location dependent data in mobile computing. *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking* (pp. 210-221).

Dunham, M.H., & Kumar, V. (1998, August). Location dependent data and its management in mobile databases. *Proceedings of the 9th Annual International Workshop on Database and Expert Systems Applications* (pp. 414-419).

Elmasri, R., & Navathe, S. B. (2004). *Fundamentals of database systems.* Boston: Addison-Wesley.

Getting, A. (1993). The global positioning systems. *IEEE Spectrum, 30*(12), 36-47.

Goodman, D. J. (1998). *Wireless personal communications systems* (Wireless Communications Series). Boston: Addison-Wesley.

Guttman, A. (1984). A dynamic index structure for spatial searching. *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data* (pp. 47-57). New York: ACM Press.

Jayaputera, J., & Taniar, D. (2005). Query processing strategies for location-dependent information services. *International Journal of Business Data Communications and Networking, 1*(2), 17-40.

Jelenkovic, P., & Radovanovic, A. (2003). Asymptotic insensitivity of least-recently-used caching to statistical dependency. *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies* (pp. 438-447).

Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A. N., & Theodoridis, Y. (2005). *{R}-Trees: Theory and applications.* Berlin; Heidelberg; New York: Springer-Verlag.

Mobile Computing Horizontal Applications Index. (2006). Retrieved from http://www.mobileinfo.com/Applications_Horizontal/index.htm

Samet, H. (1984). The Quadtree and related hierarchical data structures. *ACM Computing Surveys, 16*(2), 187-260.

Sellis, T., Roussopoulos, N., & Faloutsos, C. (1987). The R+-Tree: A dynamic index for multi-dimensional objects. *Proceedings of the 13th Very Large Data Bases Conference* (pp. 507-518).

Taniar, D., & Rahayu, J. W. (2002). A Taxonomy Of Indexing Schemes For Parallel Database Systems. *Distributed and Parallel Databases: An International Journal, 12*(1), 73-106.

Tayeb, J., Ulusoy, O., & Wolfson, O. (1998). A Quadtree-based dynamic attribute indexing method. *The Computer Journal, 41*(3), 185-200.

Theodoridis, Y., & Sellis, T. (1994, March). Optimization issues in R-Tree construction. *Proceedings of the International Workshop on Geographic Information Systems (IGIS)* (pp. 270-273).

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Research on location-dependent queries in mobile databases. *International Journal on Computer Systems: Science and Engineering, 20*(3), 77-93.

KEY TERMS

Base Station: Stationary that forwards a message from a wireless to a wired network.

Cell: A specific area served by one base station.

Location-Dependent Information Services (LDIS): A service that gives information based on the current location of mobile users.

Location-Dependent Query (LDQ): A query that includes current location.

Neighboring Cell: Adjacent cell of current cell.

Partial Global Indexing (PGI): Modification of Global Indexing technique used to include indexes of objects from a neighboring cell.

Patterns for Mobile Applications

Markus Aleksy
University of Mannheim, Germany

Martin Schader
University of Mannheim, Germany

INTRODUCTION

The increasing adoption of mobile terminals, together with the progressive development of handheld devices and the simultaneous improvement in the infrastructure of wireless communication plays a more and more important role in the development of new kinds of mobile applications. Applications that are executed on mobile terminals often require both a simultaneous interaction with other users of mobile terminals as well as with fixed or location-dependent services. Although this development is very favorable from the view of the end user, it confronts application developers with the problem that they must manage the growing complexity of the mobile applications.

A main characteristic of mobile applications is their dynamics. It is therefore important that during development of a mobile application, its dynamic adaptability—as one of the most important design criteria—must be taken into account.

In this article, we present a pattern-based structure for the development of mobile applications. After a short description of the concept of patterns, we demonstrate how these may be used in the context of the development of mobile applications.

PATTERNS

Patterns are simple and concise solutions for programming tasks frequently appearing in practice. The architect Christopher Alexander (Alexander et al., 1977; Alexander, 1979) is regarded as being the intellectual father of the patterns movement. The structure of a pattern follows certain rules and is based on the elements listed below (Gamma, Helm, Johnson, & Vlissides, 1995):

- **Pattern Name:** The name of a pattern is usually short yet descriptive and acts as an addition to the design vocabulary.
- **Problem:** Patterns should include a short description of the problem they intend to solve.
- **Solution:** The solution to the problem is described in a generally applicable way. The elements of the solution are described along with their relationships and responsibilities.
- **Consequences:** While the main consequence of using the pattern is solving the problem, there are often side effects. In order to make it easier to understand the trade-off involved in using the pattern, it is important that the potential drawbacks are explained.

Of course, this is not the only possible structure for describing a pattern. Gamma et al. (1995) also provide an extended schema, and other authors offer alternative descriptions as well (e.g., Fowler, 1997).

A pattern documents a comprehensive and good solution won from experience to a problem frequently recurring during program design. The engineer-like approach at the design of object-oriented software consists of identifying one or more patterns for a concrete programming task, which helps in coping with this task. The documentation of the patterns supports the development process, from the complete specification of the set task up to its implementation in a concrete programming language; in addition, it supplies a common language frame for the discussion of the envisaged solution between the developers.

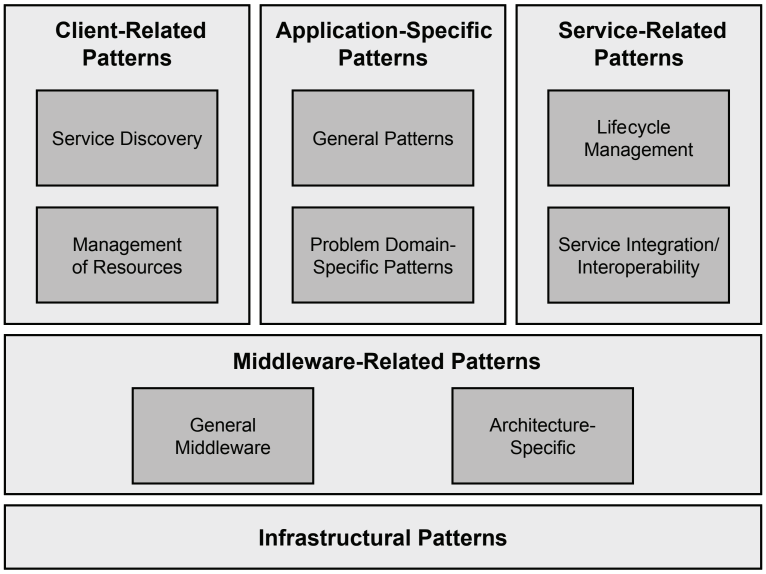
The primary benefit of a pattern lies in the description of a solution for a certain class of problems. Additional advantage arises from the fact that every pattern has a name. This simplifies the communication among software developers since one may then discuss a software structure on an abstract level. Thus, patterns are first of all programming language independent. During the design of object-oriented software, they serve as a recognized aid to structuring the outline. A detailed introduction to the concept of patterns can be found among others in Gamma et al. (1995) or Buschmann, Meunier, Rohnert, Sommerlad, and Stal (1996).

PATTERNS FOR THE DEVELOPMENT OF MOBILE APPLICATIONS

When developing a mobile application, a programmer can meanwhile resort to a variety of different patterns. One can distinguish among patterns with a different coverage (see Figure 1):

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Patterns for the development of mobile applications


Client-Related Patterns
Service Discovery
Management of Resources
Application-Specific Patterns
General Patterns
Problem Domain-Specific Patterns
Service-Related Patterns
Lifecycle Management
Service Integration/ Interoperability
Middleware-Related Patterns
General Middleware
Architecture-Specific
Infrastructural Patterns

- infrastructural patterns,
- middleware-related patterns,
- client-related patterns,
- service-related patterns, and
- application-specific patterns.

Infrastructural Patterns

Infrastructural patterns describe solutions for the administration of mobility (mobility management functions). They illustrate which elements are needed for the realization of architecture. Andrade, Logrippo, Bottomley, and Coram (2001) identify two types of patterns that fall into this category: *architectural elements* and *functional behaviors.* The patterns *home and visitor databases* as well as *security database* fall into the first category. The first pattern solves the problem of mobility between different local areas (*roaming*) of the same supplier or of different suppliers. The second is aimed at the realization of security and privacy functionality. Besides that, the authors identify broader functional behavioral patterns, such as *temporary identification, paging, authentification, ciphering,* as well as *location registration.*

Middleware Patterns

Building on the infrastructural patterns introduced above, a variety of middleware patterns can be used. Here, middleware functions as the link between different spatially distributed software components. We introduce different middleware-related patterns in the following. These were developed mainly for the realization of non-mobile applications, but it is possible to transfer them into the domain of mobile computing as well.

Patterns that fall into this category can be classified as *general design patterns,* which can be used for the realization of (almost) all types of distributed architectures, or as *architecture-specific patterns.*

General patterns, which can be used for the realization of a basic infrastructure for object-oriented systems, are described for example by Buschmann et al. (1996), Schmidt, Buschmann, Stal, Rohnert, and Sommerlad (2001), and Völter, Kircher, and Zdun (2002).

Patterns such as the *forwarder-receiver* or the *publisher-subscriber* (Buschmann et al., 1996) may be applied in a variety of applications and architectures.

On the other side, middleware patterns exist which are specific for certain architectures. For example, during the realization of a mobile client-server architecture, the patterns introduced by Völter (2001) can be used on the server side. Aarsten, Brugali, Menga, Brown, and Hirschfeld (2005) present patterns particularly suitable for the realization of three-tier client-server architectures. Another pattern falling into this category is the *broker* pattern (Buschmann et al., 1996), which gained wide popularity in the context of the common object request broker architecture standard (CORBA) defined by the Object Management Group (2004).

Patterns for the Development of Mobile Clients

During development of a mobile application, dealing efficiently with the resources of the mobile client represents an important challenge. Despite the enormous technological progress, different aspects, such as relatively small memory sizes or comparatively low network capacity, must still be taken into account.

Here, the *virtual proxy* pattern (Gamma et al., 1995) may be applied. Proxy objects are created in the context of the implementation of the mobile client. The use of this pattern has the consequence that "expensive" objects are loaded dynamically—that is, they are only loaded and instantiated if and when required. In this focus, further patterns have been identified such as the *lazy acquisition* pattern (Kircher, 2001), which aims at allocating a resource at the latest possible point in time. Furthermore, with the help of the *virtual component* pattern (Corsaro, Schmidt, Klefstad, & O'Ryan, 2002), aspects like configurability and adaptability of an application can be improved at run time.

Another factor concerns the localization of the services required by the mobile application. So that a mobile client can use a service, it must be able to identify services in a dynamic environment. The offered services can differ from one location to another. It is therefore necessary to decouple the mobile client from the services.

In order to solve the problem, for example, the *lookup* pattern (Kircher, 2001) can be applied. A central instance, the lookup service, functions as a link between mobile clients and the hard-wired services in this case. This approach is also followed in the context of different popular technologies such as CORBA (Object Management Group, 2004). Further patterns, which are related to service discovery, or more exactly a complete pattern language, are described in Pärssinen, Koponen, and Eronen (2004). In this regard, the *naming* pattern (Silva, Sousa, & Antunes, 1998) should also be mentioned.

The dynamics of the mobile clients represent another problem: they may not be "hard-wired" to the services they use. The *service abstraction layer* pattern (Vogel, 2001) can be employed to decouple these possibly heterogeneous mobile clients from the services. This pattern signifies amongst others that an additional layer should be introduced between the client application and the services. It can be realized by means of the *facade* pattern (Gamma et al., 1995). The *service abstraction layer* pattern promotes aspects like separation of concerns, generic request handling, controlled evolution, or communication transparency, for example.

This way, services tailored to the client application can be realized, which, for example depending on location, have different capabilities, such as possibly supplying various quality of service (QoS) features.

By the permanently changing business processes and the regular emergence of new technologies, the "separation of concerns" principle (Völter, 2001) should be stressed. In connection with the patterns introduced here, it may suitably be applied as a basis for the development of platform and programming language-independent mobile applications.

Service Patterns for Mobile Applications

Mobile applications, primarily if these are executed in a location-aware context, require great flexibility. The reason is that the current location of the device decides on the accessible services. Such applications must therefore possess two special qualities: the ability of ad-hoc configuration and the ability of efficiently discovering, adding, using, and administering services.

Although one could think that the aspect of the dynamic adaptability of a mobile client application concerns only the client side, this is normally not the case. The reason is that the flexible detection, use, and administration of services is not alone relevant on the client side, but also requires a flexible infrastructure on the server side. As mentioned above, the ability to efficiently administer services, that is, their discovery, addition, and use, is of essential importance. Several problems must be solved here, including the administration of no-longer-needed services or the selection of a service provider. In the following sections, we will present different patterns that are of importance for the implementation of an administrative infrastructure for the services used by a mobile client.

In the case of mobile clients, the use of the *evictor* pattern (Henning & Vinoski, 1999; Jain, 2001) appears to be the most suitable solution for the administration of services. This approach tries to cope with the aim of efficiently using services by monitoring. Every time an access to a service is carried out, a flag is set. The least recently used (LRU) services or the least frequently used (LFU) services represent candidates for release. The release of the services can happen either periodically or upon request.

It is the principal benefit of this pattern that it is transparent from the view of the client. That is, the mobile client does not have to take care of the release of the services accessed by it, at all. Moreover, the release of no-longer-needed services is ensured even in the case of technical problems, such as interruption of the communication, crash of the mobile client, and so forth. The only disadvantage of this approach is its "ignorance" of the exact time of the services' release, since the mobile client application cannot take any countermeasures in this situation. However, this disadvantage can be remedied relatively simply by providing corresponding configuration possibilities.

A further problem is the fact that every service is not always attainable. This can either be based on technical causes again or depend on the current location of the mobile client. In this context, the *leasing* pattern (Jain & Kircher, 2000) can be used. This pattern is based on a method that does not assign a service to a user for an indefinite period, but only for a particular timeframe. If the time has run out, different action variants are possible. The leasing concept—which is, for example, of essential importance in the Jini architecture (Sun Microsystems, 2003) and for which Sun Microsystems even realized an own specification (Sun Microsystems, 1999)—was until now largely neglected in mobile location-dependent applications. The work of Jain and Kircher (2000), who even use the term leasing *pattern,* clarifies that this concept does not have to be limited to Jini only. The authors elaborate the exact details, which are related to the problem definition, the structure, and possible variants of leasing. In technologies that do not directly support leasing, such as CORBA, the concept can nevertheless be realized by means of a dedicated service (Aleksy & Gitzel, 2002).

Furthermore, a number of patterns exist that promote the development of dynamically configurable software components. In connection with this, design patterns such as the *service configurator* (Kircher & Jain, 1997) or the *component configurator* (Schmidt et al., 2001) can be mentioned. Additional patterns that can be used for the dynamic administration of resources are, among others, given in Welch, Marinucci, Masters, and Werme, 2002).

Application-Specific Patterns

Two principally different pattern types can be discriminated in this category: *general patterns* as well as *application-specific patterns.* The first category is domain independent—that is, it can be used in a large variety of applications. This kind of patterns is described, for example, in Gamma et al. (1995).

On the other side, a variety of patterns exists, which very narrowly are connected to a concrete problem domain. In location-dependent mobile applications, the context is of essential importance since it is the determining factor for the range of utilizable services. Dockhorn, Ferreira Pires, and van Sinderen (2005) describe three architectural patterns that can be employed especially in the development of context-specific applications. The *event control action* pattern serves for the decoupling of contexts and the actions being based on them. The use of this pattern promotes the "separation of concerns" principle. The *context sources and manager hierarchy* pattern describes a hierarchical architecture for the processing of context information. The *action patterns* pattern aims at providing a structural schema to enable coordination of actions as well as decoupling of action implementations from action purposes.

FUTURE TRENDS

Although a variety of patterns that can be used in the area of mobile applications was already identified, that process is not completed yet. The increasing popularity of mobile applications is the reason that new applications or types of applications emerge, which in turn can lead to the discovery of new patterns. This development is not surprising since the pattern concept builds on the reuse of solutions already tested successfully. Therefore, it has to be expected that, in the future, the already-established patterns will be continuously complemented by newly discovered patterns. In this regard, the aspect of the administration and use of the location context is primarily of special importance.

CONCLUSION

In this article, we have introduced different patterns that can be used for the development of mobile applications. These cover different aspects such as the realization of a basic infrastructure, middleware specifics, the development of mobile client applications, as well as patterns supporting the server side of mobile applications.

This way, using existing patterns, a complete basic infrastructure can be realized for the architecture of mobile applications. Based on the patterns introduced here, mobile applications can, in general, be developed faster and more economically, and also their quality may be enhanced by the adoption of "best practice" techniques already tested.

Nevertheless, the domain of mobile applications requires further investigation to identify patterns, which are new and not discovered until now.

REFERENCES

Aarsten, A., Brugali, D., Menga, G., Brown, K., & Hirschfeld, R. (2005). *Patterns of three-tier client-server architectures.* Retrieved from http://members.aol.com/kgb1001001/Articles/threetier/threetier.htm

Aleksy, M., & Gitzel, R. (2002). Design and implementation of a leasing service for CORBA-based applications. *Proceedings of the 1st International Symposium on Cyber Worlds: Theories and Practices (CW 2002)*, Tokyo, Japan, (pp. 54-61).

Alexander, C. (1979). *The timeless way of building.* New York: Oxford University Press.

Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., & Angel, S. (1977). *A pattern language.* New York: Oxford University Press.

Andrade, R., Logrippo, L., Bottomley, M., & Coram, T. (2001). A pattern language for mobility management. *Proceedings of the 6th European Conference on Pattern Languages of Programs (EuroPLoP 2001)*, Irsee, Germany.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., & Stal, M. (1996). *Pattern-oriented software architecture—A system of patterns.* Chichester, UK: John Wiley & Sons.

Corsaro, A., Schmidt, D. C., Klefstad, R., & O'Ryan, C. (2002). Virtual component: A design pattern for memory-constrained embedded applications. *Proceedings of the 9th Conference on Pattern Language of Programs (PLoP 2002)*, Monticello, IL.

Dockhorn C., Ferreira Pires, L., & van Sinderen, M. (2005, May). Architectural patterns for context-aware services platforms. *Proceedings of the 2nd International Workshop on Ubiquitous Computing (IWUC 2005)* (pp. 3-18). Miami, FL.

Fowler, M. (1997). *Analysis patterns: Reusable object models.* Boston: Addison-Wesley.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* Boston: Addison-Wesley.

Henning, M., & Vinoski, S. (1999). *Advanced CORBA programming with C++.* Boston: Addison-Wesley.

Jain, P. (2001). Evictor. *Proceedings of the 8th Patterns Languages of Programs Conference (PLoP 2001)*, Monticello, IL.

Jain, P., & Kircher, M. (2000). *Leasing. Proceedings of the 7th Patterns Languages of Programs Conference (PLoP 2000)*, Monticello, IL.

Kircher, M. (2001). Lazy acquisition. *Proceedings of the 6th European Conference on Pattern Languages of Programs (EuroPLoP 2001)*, Irsee, Germany.

Kircher, M., & Jain, P. (2000). Lookup. *Proceedings of the 5th European Conference on Pattern Languages of Programs (EuroPLoP 2000)*, Irsee, Germany.

Kircher, M., & Jain, P. (1997). Service configurator: A pattern for dynamic configuration of services. *Proceedings of the 3rd Conference on Object-Oriented Technologies and Systems,* USENIX.

Object Management Group. (2004). *The common object request broker: Architecture and specification. Version 3.0.3.* OMG Technical Document Number formal/04-03-12. Retrieved from ftp://ftp.omg.org/pub/docs/formal/04-03-12.pdf

Pärssinen, J., Koponen, T., & Eronen, P. (2004). Pattern language for service discovery. *Proceedings of the 9th European Conference on Pattern Languages of Programs (EuroPLoP 2004)*, Irsee, Germany.

Schmidt, D, Buschmann, F., Stal, M., Rohnert, H., & Sommerlad, P. (2001). *Pattern-oriented software architecture—Patterns for concurrent and networked objects.* Chichester: John Wiley & Sons.

Silva, A. R., Sousa, P., & Antunes, M. (1998, August). Naming: Design pattern and framework. *IEEE Proceedings of the 22nd Annual International Computer Software and Applications Conference (COMPSAC 98)* (p. 316). Vienna, Austria.

Sun Microsystems. (1999). *Distributed leasing specification.* Retrieved from http://wwws.sun.com/software/jini/specs/jini10specs/lease-spec.html

Sun Microsystems. (2003). *JiniTM architecture specification—Version 2.0.* Retrieved from http://wwws.sun.com/software/jini/specs/jini2_0.pdf

Völter, M. (2001). Server-side components—a pattern language. *Proceedings of the 6th European Conference on Pattern Languages of Programs (EuroPLoP 2001)*, Irsee, Germany.

Völter, M., Kircher, M., & Zdun, U. (2002). Object-oriented remoting—Basic infrastructure patterns. *Proceedings of the 1st Nordic Conference on Pattern Languages of Programs (VikingPLoP 2002)*, Højstrupgård, Denmark.

Vogel, O. (2001). Service abstraction layer. *Proceedings of the 6th European Conference on Pattern Languages of Programs (EuroPLoP 2001)*, Irsee, Germany.

Welch, L. R., Marinucci, T., Masters, M. W., & Werme, P. V. (2002). Dynamic resource management architecture patterns. *Proceedings of the 9th Conference on Pattern Language of Programs (PLoP 2002)*, Monticello, IL.

KEY TERMS

Client: A computational context that invokes remote operations on a service.

Middleware: The link between different locally distributed software components.

Pattern: A simple and concise solution to programming tasks frequently appearing.

Pattern Language: A collection of related patterns in a certain domain.

Separation of Concerns: A key concept for the development of software systems is the separation of different areas of responsibility. Different aspects of a software system are handled independently of each other and integrated at a later time.

Service: Offers a certain functionality that can be requested by different service users**.** The type of functionality offered by the service is not relevant and should be regarded as being opaque.

Service Provider: Realizes a certain service at least partially.

Peer-to-Peer Cooperative Caching in Mobile Environments

Chi-Yin Chow
University of Minnesota – Twin Cities, USA

Hong Va Leong
The Hong Kong Polytechnic University, Hong Kong

Alvin T. S. Chan
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

An infrastructure-based mobile environment is formed with a wireless network connecting mobile hosts (MHs) and mobile support stations (MSSs). MHs are clients equipped with portable devices, such as laptops, personal digital assistants, cellular phones, and so on, while MSSs are stationary servers providing information access for the MHs residing in their service areas. With the recent widespread deployment of contemporary peer-to-peer (known as P2P throughout this chapter) wireless communication technologies, such as IEEE 802.11 (IEEE Standard 802-11, 1997) and Bluetooth (Bluetooth SIG, 2004), coupled with the fact that the computation power and storage capacity of most portable devices have been improving at a fast pace, a new information sharing paradigm known as P2P information access has rapidly taken shape. The MHs can share information among themselves rather than having to rely solely on their connections to the MSS. This article reviews a hybrid communication framework—that is, mobile cooperative caching—which combines the P2P information access paradigm into the infrastructure-based mobile environment.

BACKGROUND

In mobile environments, there are two different types of communication architecture, *infrastructure* and *ad hoc based.* The infrastructure-based mobile communication architecture is formed with MHs and MSSs. The MHs can only retrieve their desired data items from MSSs, either by requesting them over shared point-to-point channels (*pull-based data dissemination model*) or catching them from scalable broadcast channels (*push-based data dissemination model*) or through the utilization of both types of channels (*hybrid data dissemination model*). This type of communication architecture is the most commonly deployed one in real life.

The emergence of the state-of-the-art P2P communication technologies leads to the development of an ad-hoc-based mobile communication architecture, also known as a *mobile ad-hoc network* (MANET). In MANETs, the MHs can share information among themselves without any help of MSSs. This kind of sharing paradigm is also referred to as a P2P *data dissemination model.*

In a pull-based environment, the MHs have to retrieve their desired data items from the MSS whenever they encounter local cache misses. Since the mobile environment is characterized by limited bandwidth, the communication channel between the MSS and the MHs would potentially become a scalability bottleneck in the system, as it serves an enormous number of MHs. Although push-based and hybrid data dissemination models are scalable, the MHs adopting these two models generally suffer from longer access latency and higher power consumption than those adopting the pull-based one, as they need to tune in to the broadcast channel and wait for the broadcast channel index or their desired data items to appear. Furthermore, since the data items are broadcast sequentially, the MHs experience longer access latency with an increasing number of data items being broadcast.

MANET is practical to a mobile system with no fixed infrastructure support, such as battlefield, rescue operations, and so on (Fife & Gruenwald, 2003). However, it is not suitable for commercial mobile applications. In MANETs, the MHs can rove freely and disconnect themselves from the network at any instant. These two particular characteristics lead to dynamic changes in the network topology. As a result, the MHs could suffer from long access latency or access failure when the peers holding the desired data items are far way or unreachable. The latter situation is caused by network partitioning (Wang & Li, 2002) or client disconnection.

The inherent shortcomings of the infrastructure- and ad-hoc-based communication architecture lead to a result that a mobile application adopting either one of these architectures alone would not be as appropriate in most real commercial

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

settings. In reality, long access latency or access failure could possibly cause the abortion of valuable transactions or the suspension of critical activities, so that it is likely to reduce user satisfaction and loyalty, and potentially bring damages to the organization involved. The drawbacks of the existing mobile data dissemination models motivate researchers to develop a novel hybrid communication framework—mobile cooperative caching—in which a convectional infrastructure-based mobile communication framework is used in combination with a P2P data dissemination paradigm for deploying mobile information access applications in reality.

MOBILE COOPERATIVE CACHING

Recently, mobile cooperative caching has been drawing increasing attention. Several mobile cooperative caching schemes were proposed during the preceding years. These works can be divided into two major categories: *cooperative data dissemination* and *cooperative cache management.* The work of cooperative data dissemination (Lau, Kumar, & Venkatesh, 2002; Papadopouli & Schulzrinne, 2001; Sailhan & Issarny, 2003; Shen, Das, Kumar, & Wang, 2004) mainly focuses on designing protocols for the MHs to search their desired data items and forward the data items from source MHs or MSSs to them in a mobile environment. The work pertaining to cooperative cache management focuses on designing protocols and algorithms for the MHs to manage their cache space, not only with respect to themselves, but also with respect to their peers, in order to improve system performance along such design dimensions as *cooperative data replica allocation* (Hara, 2001, 2002a, 2002b; Hara, Loh, & Nishio, 2003), *cooperative cache invalidation* (Hayashi, Hara, & Nishio, 2003), and *cooperative cache admission control* and *cache replacement* (Lim, Lee, Cao, & Das, 2003; Chow, Leong, & Chan, 2004, 2005).

COOPERATIVE DATA DISSEMINATION

Sailhan and Issarny (2003) propose an intuitive cooperative data dissemination scheme for a MANET environment. If an MH can directly connect to an MSS, it would obtain the required data items from the MSS; otherwise, the MH has to enlist its peers at a distance less than the MSS for help to turn in the required data items. If no such peer caches the data items, the peers route the request to the nearest MSS. A local cache replacement strategy is also proposed for the MH based on the access probability and time-to-live of the cached data items, and the estimated energy cost of retrieving them.

A similar cooperative data dissemination scheme is designed to support continuous media access in MANETs (Lau et al., 2002). Two data location schemes, namely *cache-state* and *reactive,* are proposed for the MHs to determine the nearest data source that can be either the cache of their peers or the original servers to retrieve their desired multimedia objects. Cache-state is a proactive scheme, whereas reactive is an on-demand scheme. The performance evaluation result shows that the reactive scheme outperforms the cache-state one in terms of network traffic, quality of service (QoS), and access latency.

7DS (seven degrees of separation) (Papadopouli & Schulzrinne, 2001) is another cooperative data dissemination scheme that is used as a complementary component to the infrastructure support with power conservation. When an MH fails to connect to the MSS to retrieve its desired data items, it would attempt to search its neighboring 7DS peers for them. The power conservation scheme adjusts the MHs' degree of activity or participation in 7DS based on their available battery levels.

Shen et al. (2004) propose another cooperative data dissemination scheme with power conservation, called *energy-efficient cooperative caching with optimal radius* (ECOR), in a mobile environment. In ECOR, an optimal radius (in number of hops) is estimated by an analytical model that considers the MH's location, data access probability, and network density for each data item. The MHs exchange the cache content and the optimal radius of each cached data item among themselves. When an MH encounters a local cache miss, if it finds that any peers cache its desired data item, and the distance between the MH and the peer is within the optimal radius based on its local state, the MH sends a request message to the peer that is the closest to the selected holder of the data item. Otherwise, the MH obtains the data items from the MSS.

Yin and Cao (2004) propose three other cooperative caching schemes, called *CacheData, CachePath,* and *HybridCache.* The idea of CacheData is that an MH caches a passing-by data item, if the data item is popular and a condition that all requests for the data items are not originated by the same MH is satisfied. For CachePath, the MH caches path information of the passing-by data item instead of the data item. To conserve cache space, an MH does not cache path information of all passing-by data items. It only caches the path information of a data item if it is closer to the requesting MH than the MSS. HybridCache is a hybrid scheme that combines both CacheData and CachePath. An MH either applies CacheData or CachePath based on three factors: data item size, data item time-to-live, and the distance between the MH's distance to the data holder and the distance to the MSS.

COOPERATIVE CACHE MANAGEMENT

All literature related to cooperative cache management can be further divided into four sub-categories: *cooperative* *data replica allocation, cooperative cache invalidation, cooperative cache admission control,* and *cooperative cache replacement.*

Cooperative Data Replica Allocation

Data replica allocation techniques (Hara, 2001, 2002a, 2002b; Hara et al., 2003) are adopted in mobile cooperative caching to improve data accessibility, in order to alleviate the network partitioning problem. Hara (2001) proposes three data replica allocation schemes: *SAF* (static access frequency), *DAFN* (dynamic access frequency and neighborhood), and *DCG* (dynamic connectivity-based group). The MHs applying SAF only consider their own individual access probability to each data item. DAFN extends SAF to take the access probability to each data item of the MHs' connected neighborhoods into account. Finally, DCG groups the MHs with highly stable connection together. A group of MHs possesses high connection stability, as they form a *biconnected component* in the network. DCG considers the access probability to each data item of all MHs in the same group. The performance evaluation result shows that DCG gives the highest data accessibility, but it incurs higher network traffic than the other two schemes. Thus, DCG can be considered as a scheme that trades network traffic for data accessibility.

These three data replica allocation schemes are then adapted to a push-based mobile environment (Hara, 2002a). Other than data access probability, the schemes also consider the latency on accessing data items from the peers and broadcast channel. Furthermore, the proposed replica allocation schemes are further extended, namely, Extended SAF (*E-SAF*), Extended DAFN (*E-DAFN*), and extended DCG (*E-DCG*), to consider periodic data update by allowing the extended allocation schemes to consider the remaining time until next update of each data item (Hara, 2002b). In addition to the access probability, Hara et al. (2003) also consider the stability of radio links. The stability of a radio link is defined as the remaining time period that two MHs will still be connected to each other. The longer the time period indicates the higher the stability of a ratio link.

Huang, Chen, and Peng (2003) propose another distributed data replica allocation scheme in MANETs, called *DRAM,* to improve data accessibility and reduce network traffic pretending to the replication mechanism. DRAM extends E-DCG (Hara, 2002b) to consider a group mobility pattern for data replica allocation. It is assumed that some MHs tend to roam together and they share a common access range. To discover the group mobility pattern among MHs, a distributed clustering algorithm is adopted to cluster several MHs who possess a similar mobility pattern into a group. The clustering algorithm is executed periodically to adapt to the changes in network topology. Then, the data replicas are allocated to each group member based on group access probability to the data items and the remaining time until the next update on them. DRAM is found to perform better than E-DCG in terms of data accessibility and network traffic.

Cooperative Cache Invalidation

Hayashi et al. (2003) propose two cache invalidation schemes, namely *update broadcast* and *connection rebroadcast.* The former is a straightforward, flooding-based scheme. An MH that caches an original copy of a data item broadcasts an invalidation report to other peers when that MH updates the data item. The latter can be referred to as a cooperative cache invalidation scheme. When two MHs are newly connected to each other, they broadcast their collected cache invalidation information to their connected peers. The newly connected MHs and other peers receiving their broadcast information update their own previously received cache invalidation information to identify any obsolete data items in their cache. The performance evaluation result shows that the connection rebroadcast scheme reduces the number of accesses to invalid cached data items, but it incurs higher network traffic than update broadcast scheme.

Cooperative Cache Admission Control and Cache Replacement

Lim et al. (2003) propose a cooperative caching scheme for Internet-based MANETs, namely IMANET. In IMANET, a simple, flooding-based searching scheme is proposed for the MHs to search their desired data items in the network. IMANET also provides two cooperative data management protocols: *cooperative cache admission control* and *cache replacement.* For the cooperative cache admission control protocol, an MH determines whether to cache a data item based on the distance between itself and the data source that can be either other peers caching the data item or the MSS. For the cooperative cache replacement protocol, called *time and distance sensitive* (TDS), a victim data item is selected to be evicted from the cache by an MH based on two factors: the distance between itself and other peers caching the victim or the MSS, and the freshness of the distance information. The distance information is updated when the corresponding data item is accessed by other MHs. Since the network topology changes frequently, the distance information could become outdated, as it has not been updated for a long time.

There is a need for cooperating peers to cache useful data items together, so as to improve cache hit from peers. This could be realized by capturing the data requirement of individual peers in conjunction with their mobility patterns. Two group-based mobile cooperative caching schemes, namely, GroCoca (Chow et al., 2004) and DGCoca (Chow et al., 2005), make use of a concept of a *tightly coupled group* (TCG), which is defined as a group of MHs that are *geographically* and *operationally close*—that is, sharing common mobility and data access patterns. Two MHs are considered to be geographically and operationally close based upon their locations and the set of data items they access respectively.

GroCoca is a *centralized* group-based mobile cooperative caching scheme, in which the MSS uses an incremental clustering algorithm to discover TCGs based on the weighted average distance and data access similarity of any two MHs. In GroCoca, when an MH encounters a local cache miss and its peer can turn in its desired data item to it, it only caches the data item if its local cache has not been fully occupied or the peer is not belonging to its TCG. In other words, the MHs do not cache the data items that are provided by their TCG members, on the belief that the data items can be readily available from the peer if needed.

On the contrary, DGCoca is a *distributed* group-based mobile cooperative caching scheme, in which a stable neighbor discovery algorithm is proposed for the MHs to discovery their own TCG's members dynamically without any help of the MSS. The MHs adopting DGCoca not only make use of the cooperative cache admission control protocol proposed in GroCoca, but they also perform cooperative cache replacement to further improve data accessibility. The proposed cooperative cache replacement protocol possesses three important properties. First, the most valuable data items are always retained in the local cache. Second, in a local cache, a data item which has not been accessed for a long period will be replaced eventually. Third, in a TCG, a data item which "spawns" replica is first replaced in order to increase the effective cache size.

FUTURE TRENDS

Mobile cooperative caching is one of the most promising techniques to improve system performance in mobile environments. The future mobile cooperative caching scheme will focus on investigating security and privacy issues. Since an MH can cache outdated, counterfeit, or harmful data items, the access to these data items potentially brings damages to other peers. Therefore, it is necessary to design a mechanism for an MH to ensure the freshness, reliability, and safety of the data items retrieved from other peers. Also, when an MH accesses cached information from others, the MH may tell some personal information about them, for example, personal data, preferences, location information, data access history, and so forth. To resolve this problem, there should be a permission control mechanism on accessing a cached data item, so that the MHs can choose what kinds of information not to be disclosed. Also, an anonymous data sharing protocol should be developed to protect the privacy of the participating MHs in a mobile cooperative caching environment because their personal information may be revealed by other peers through their requests.

CONCLUSION

Mobile cooperative caching is a novel hybrid communication architecture that combines the P2P data dissemination model into a conventional infrastructure-based communication architecture. In mobile cooperative caching, the MHs can retrieve their desired data items not only from the MSS, but also from their peers. The MHs also manage their cache space with respect to themselves and their peers to improve system performance, such as cooperative data replica allocation, cooperative cache invalidation, cooperative cache admission control, and cache replacement. The future trend of mobile cooperative caching will be focusing on how to resolve the security and privacy issues.

REFERENCES

Bluetooth SIG. (2004). *Bluetooth specification v2.0.* Retrieved from http://www.bluetooth.org

Chow, C.-Y., Leong, H. V., & Chan, A. T. S. (2004). Group-based cooperative cache management for mobile clients in a mobile environment. *Proceedings of the 33rd International Conference on Parallel Processing (ICPP)* (pp. 83-90). Montreal, Canada.

Chow, C.-Y., Leong, H. V., & Chan, A. T. S. (2005). Distributed group-based cooperative caching in a mobile broadcast environment. *Proceedings of the 6th International Conference on Mobile Data Management (MDM)* (pp. 97-106). Ayia Napa, Cyprus.

Fife, L. D., & Gruenwald, L. (2003). Research issues for data communication in mobile ad-hoc network database systems. *ACM SIGMOD Record*, *32*(2), 42-47.

Hara, T. (2001). Effective replica allocation in ad hoc networks for improving data accessibility. *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)* (pp. 1568-1576). Anchorage, AK.

Hara, T. (2002). Cooperative caching by mobile clients in push-based information systems. *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM)* (pp. 186-193).McLean, VA.

Hara, T. (2002). Replica allocation in ad hoc networks with periodic data update. *Proceedings of the 3rd International Conference on Mobile Data Management (MDM)* (pp. 79-86).Singapore.

Hara, T., Loh, Y.-H., & Nishio, S. (2003). Data replication methods based on the stability of radio links in ad hoc networks. *Proceedings of the 6th International Workshop on*

Mobility in Databases and Distributed Systems, in conjunction with the 14th International Conference on Database and Expert Systems Applications (DEXA) (pp. 969-973). Prague, Czech Republic.

Hayashi, H., Hara, T., & Nishio, S. (2003). Cache invalidation for updated data in ad hoc networks. *Proceedings of the 11th International Conference on Cooperative Information Systems (CoopIS)* (pp. 516-535). Catania, Italy.

Huang, J.-L., Chen, M.-S., & Peng, W.-C. (2003). Exploring group mobility for replica data allocation in a mobile environment. *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM)* (pp. 161-168). New Orleans, LA.

IEEE Standard 802-11. (1997). *IEEE standard for wireless LAN medium access control (MAC) and physical layer (PHY) specification*.

Lau, W. H. O., Kumar, M., & Venkatesh, S. (2002). A cooperative cache architecture in support of caching multimedia objects in MANETs. *Proceedings of the 5th ACM International Workshop on Wireless Mobile Multimedia, in conjunction with the 8th International Conference on Mobile Computing and Networking (MobiCom)* (pp. 56-63). Atlanta, GA.

Lim, S., Lee, W.-C., Cao, G., & Das, C. R. (2003). A novel caching scheme for Internet based mobile ad hoc networks. *Proceedings of the 12th IEEE International Conference on Computer Communications and Networks (ICCCN)* (pp. 38-43). Dallas, TX.

Papadopouli, M., & Schulzrinne, H. (2001). Effects of power conservation, wireless coverage and cooperation on data dissemination among mobile devices. *Proceedings of the 2nd ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)* (pp. 117-127). Long Beach, CA.

Sailhan, F., & Issarny, V. (2003). Cooperative caching in ad hoc networks. *Proceedings of the 4th International Conference on Mobile Data Management (MDM)* (pp. 13-28). Melbourne, Australia.

Shen, H., Das, S. K., Kumar, M., & Wang, Z. (2004). Cooperative caching with optimal radius in hybrid wireless network. *Proceedings of the 3rd International IFIP-TC6 Networking Conference* (pp. 841-853). Athens, Greece.

Wang, K. H., & Li, B. (2002). Efficient and guaranteed service coverage in partitionable mobile ad-hoc networks. *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)* (pp. 1089-1098). New York.

Yin, L., & Cao, G. (2004). Supporting cooperative caching in ad hoc networks. *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)* (pp. 2537-2547). Hong Kong.

KEY TERMS

Biconnected Component: In graph theory, a biconnected component is a maximal subset of edges of a connected graph such that the corresponding induced subgraph cannot be disconnected by deleting any one vertex.

Cache Admission Control: A set of policies for a mobile client to decide on whether to cache a data item.

Cache Invalidation: A mechanism for a mobile client to check whether its cached data items have been updated by someone else.

Cache Replacement: When a cache has no room for storing a data item, the least valuable cached data item is literately removed from the cache until there is enough space for caching the required data item.

Data Replica Allocation: An algorithm for allocating some selected data items to a single mobile client or a group of mobile clients based on some prescribed criteria.

Mobile Ad-Hoc Network (MANET): A peer-to-peer mobile communication architecture, in which the mobile clients can share information among themselves without any help of mobile support stations.

Network Partition: There is no mobile client acting as a gateway between two groups of mobile clients in a mobile environment.

Tightly Coupled Group: A group of mobile clients share common mobility pattern and data affinity.

Time-to-Live: The remaining time period of a data item to be updated or evicted from a cache.

Pen-Based Mobile Computing

Bernie Garret
University of British Columbia, Canada

INTRODUCTION

The original idea of a portable computer is credited to Alan Kay of the Xerox Palo Alto Research Center who suggested the idea in the 1970s (Kay, 1972a, 1972b; Kay & Goldberg, 1977). He envisioned a notebook-sized portable computer named the "Dynabook" that could be used for all of the user's information needs and using wireless network capabilities for connectivity.

BACKGROUND

Origins: Laptop Computers

The first actual portable "laptop" computers appeared in 1979: the Grid Compass Computer was designed in 1979 by William Moggridge for Grid Systems Corporation (Stanford University, 2003). The Grid Compass was one-fifth the weight of any model equivalent in performance and was used by NASA on the space shuttle program in the early 1980s. Portable computers continued to develop in the 1980s onwards, and most weighed around about 5 kg without any peripherals.

In 1984, Apple Computer introduced its Apple IIc model (Abbate, 1999), a true notebook-sized computer weighing about 5 kg without a monitor. The Apple IIc had an optional LCD panel monitor which made it genuinely portable and was therefore highly successful.

In 1986, IBM introduced its IBM Convertible PC with 256KB of memory; it was also a commercial success. By many, this is considered the first true laptop (mainly due to its clamshell design) that was shortly copied by other manufacturers such as Toshiba who were also successful with IBM laptop clones (Allen, 2001; Cringely, 1996). These devices retained the A4 size footprint, full QWERTY keyboards, and weighed between 3 and 4 kg (IBM, 2006). Following these innovations "tablet" PCs with a flat A4 footprint and a pen-based interface began to emerge in the 1990s.

There were several devices in the 1970s that explored the tablet, but in 1989 the Grid Systems GRiDPad was released, which was the world's first IBM PC-compatible tablet PC that featured handwriting recognition as well as a pen-based point-and-select system. In 1992, Microsoft released Microsoft Windows for Pen Computing, which had an application programming interface (API) that developers could use to create pen-enabled applications. Focusing specifically on devices that use the pen as the primary input device, this interface has been most successfully adopted in the new breed of small, highly portable personal digital assistants (PDAs).

Personal Digital Assistants

In 1984 David Potter and his partners at PSION launched the "PSION Organiser" which retailed for just under £100 (Troni & Lowber, 2001). It was a battery-powered, 14 x 9cm, block-shaped unit with an alphabetic keyboard and small LCD screen, with 2K of RAM, 4KB of applications in ROM, and a free 8KB data card (which had to be reformatted using ultraviolet light for reuse). Compared to the much larger notebook computers of the time, it was a revolutionary device, but because of its more limited screen size and memory, it fulfilled a different niche in the market and began to be used for personal information management and stock inventory purposes (with a plug-in barcode reader).

In the late 1980s and throughout the 1990s, PSION continued to develop commercially successful small computing devices incorporating a larger LCD screen, and a new fully multi-tasking graphical user interface (before even Microsoft had got Windows up and running). These small devices were truly handheld. The PSION 3c (launched in 1991) dimensions were 165 x 85 x 22 mm, with a 480 x 160 pixel LCD screen, and the device weighed less than 400 g. A small keyboard and innovative touch-sensitive pad provided control of the cursor, and graphical icons could be selected to start applications/functions and select items from menus. The small keyboard proved difficult to use however, and the following 5c model in 1997 used an innovative foldout miniature QWERTY keyboard. These genuinely "handheld" devices with their interface innovations and ability to synchronize data with a host personal computer made the PSION models particularly successful and firmly established the personal digital assistant as a portable computing tool for professionals.

Pen-Based Interfaces for the PDA

The limitations of keyboard-based data entry for handheld devices had been recognized, and following PSION's lead, Apple Computers introduced the Newton Message Pad in 1993. This device was the first to incorporate a touch-sensitive

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

screen with a pen-based graphical interface and handwriting-recognition software. Although moderately successful the device's handwriting recognition proved slow and unreliable, and in 1998 Apple discontinued its PDA development. However, the PDA market was now becoming firmly based upon devices using pen-based handwriting recognition for text entry, and in mid-2001, PSION, with dwindling sales and difficulties with business partnerships, ceased trading. US Robotics launched the "Palm Pilot" in 1996 using its simple "Graffiti" handwriting recognition system, and Compaq released the "iPAQ" in 1997 incorporating the new Microsoft "Windows CE/Pocket PC" operating system with the first PDA color screen.

Microsoft's relatively late entry into this market reflected the considerable research and development it undertook into developing a user-friendly pocket PC handwriting recognition interface. This remains a highly competitive field, and from November 2002 PalmSource (the new company owning the Palm Operating System) replaced the Graffiti system with Computer Intelligence Corporation's JOT as the standard and only handwriting software on all new Palm-powered devices. Computer Intelligence Corporation (CIC) was founded in conjunction with the Stanford Research Institute (SRI) based on research conducted by SRI on proprietary pattern recognition technologies (CIC, 1999). The original Graffiti system relied on the user learning a series of special characters, which while simple was irksome to many users. The CIC JOT and Microsoft Pocket PC systems have been developed to avoid the use of special symbols or characters and allow the user to input more naturally by using standard upper and lowercase printed letters. Both systems also recognize most of the original Palm Graffiti-based special characters. In 2006 Palm introduced the Windows Mobile (Pocket PC) operating system on its own high-end devices.

The Thumb Board Text Interface

The arrival of the short messaging service (SMS), otherwise known as text messaging for cellular phones, in the late 1990s led several PDA manufacturers to adopt an alternative Thumb Board interface for their PDAs. SMS allows an individual to send short text and numeric messages (up to 160 characters) to and from digital cell phones and public SMS messaging gateways on the Internet. With the widespread adoption of SMS by the younger generation, thumb-based text entry (using only one thumb to input data on cell phone keypads) became popular (Karuturi, 2003). Abbreviations such as "C U L8er" for "See you later" and "emoticons" or "smileys" to reduce the terseness of the medium and give shorthand emotional indicators developed. The rapid commercial success of this input interface inspired the implementation of Thumb Board "keyboards" on some PDAs (such as the Palm Treo 600) for text interface. Clip-on Thumb Board input accessories have also been developed for a range of PDAs.

Tablet Format PCs

The tablet PC provides a small (usually 10 x 12" screen size) rectangular format device equipped with a sensitive screen designed to interact with a device-specific pen. The pen is used directly to write or tap on the screen. It can be used in place of a keyboard or mouse for data entry; to select, drag, and open files; to draw on the screen; and to handwrite notes and communications. Tablet PCs also incorporate handwriting recognition and conversion to text software. Unlike a touch-sensitive screen, the Tablet PC screen only receives information from the device-specific pen. It will not take information from pressure applied to the screen, so users can rest their hands on the screen and write in a more natural way. Most Tablet PCs also come with optional attachable keyboards and docking stations so they can be used in the same way as a desktop computer.

A pen-based interface for the PC was developed in the early 1990s and was originally envisaged as a challenge to the mouse. Microsoft launched "Pen Extensions for Windows 3.1" in 1991 calling it "Windows for Pen Computing." The system was designed to use plug-in slate and pen systems. However, pen-based systems would take another 10 years to become established. Shortly after its launch a number of companies introduced hardware to support it. Among them were Samsung, Fujitsu, Compaq, Toshiba, and IBM. The original IBM ThinkPad was designed as a pen-based computer. However, these pen-based systems were not well received, as many users found the Windows interface difficult to use with the stylus, and by 1995 sales of pen-based systems failed to support their further mainstream development. Bill Gates remained a strong supporter of the interface, and Microsoft decided to reintroduce pen computers as the "Tablet PC" in 2002. This time the Tablet PC specification was more successful as the use of touch-screen technologies for the pen (not well developed in the 1990s), handwriting recognition, and better integrated smaller devices made the portable tablet more acceptable for consumers.

The tablet PC has proved popular for specialist uses such as in the classroom, for creative artistic use, or more recently as the platform of choice for electronic flight planning/mapping software in aviation. A growing number of manufacturers are now producing Tablet PC hardware. However, the format still retains a far smaller proportion of the mobile PC market compared to laptops and PDAs.

MULTIMEDIA AND WIRELESS INTEGRATION

Current developments in pen-based computer interfaces are exploring the use of multimedia, voice recognition, and wireless connectivity. The expansion of memory capabilities and processor speeds for mobile computing devices has enabled audio recording, digital music storage/playback, and now digital image and video recording/playback to be integrated into these devices. This and the integration of wireless network and cellular phone technologies have expanded their utility considerably.

One of the mobile computer user. Audio is attractive for mobile applications because it can be used when the user's hands and tablet interfaces remains the output display, it can be used in conditions of low screen visibility, and it may consume less power than text-based input in the PDA. The latest PDA interface innovations include voice command and dictation recognition (voice to text), voice dialing, image-based dialing (for cell phone use, where the user states a name or selects an image to initiate a call), audio memo recording, and multimedia messaging (MMS). Several devices (e.g., the new Carrier Technologies I-Mate and Palm Treo) also incorporate a digital camera.

Wireless connectivity has enabled Internet connectivity, enabling users to access e-mail, text/graphical messaging services (SMS and MMS), and the Web remotely. These developments are gradually expanding the PDA's functionality into a true multi-purpose tool.

FUTURE TRENDS

One of the key limitations of PDA and tablet interfaces remains the output display screen size, brightness, and resolution. Issues of resolution and brightness continue to hinder many potential applications for this technology. As input technologies improve, and voice and handwriting recognition come of age, then attention to the display capabilities of these devices will need to be addressed before their full potential can be realized.

Coding PDA applications to recognize handwriting, speech, and incorporate multimedia requires additional code beyond traditionally coded interfaces. PDA application design and development environments need to support this functionality more effectively in order to promote the development of more complex mobile applications.

Data and device security are key areas for highly portable networked PDAs, and the first viruses for PDAs have started to emerge (Melnick, Dinman, & Muratov, 2004; BitDefender 2004). As multimedia interfaces develop, the specific security issues that they entail (such as individual voice recognition and prevention of data corruption of new file formats) will also need to be addressed.

CONCLUSION

Since the early models, manufacturers have continued to introduce smaller and improved portable computers, culminating in the latest generation of powerful handheld PDAs offering fast (400 MHz and faster) processors, with considerable memory (64MB of ROM and 1GB of RAM or more). This area of technological development remains highly competitive, and by necessity, the user interface for these devices has developed to fulfill the portable design brief, including the use of pen- and voice-based data input, collapsible LCD displays, wireless network connectivity, and now cell phone integration. Modern PDAs are much more sophisticated, lightweight devices and are arguably much closer to Kay's original vision of mobile computing than the current laptop or tablet computers, and possibly have the potential to replace this format with future interface developments. Indeed, if the interface issues are successfully addressed, then it is probable that these devices will outsell PCs in the future and become the major computing platform for personal use.

REFERENCES

Abbate, J. (1999). Getting small: A short history of the personal computer. *Proceedings of the IEEE, 87*(9), 1695-1698.

Allen, R.A. (2001). *A history of the personal computer: The people and the technology III* (pp. 11-20). London; Ontario, Canada: Allen Publishing.

BBC. (2004). *First pocket PC virus discovered.* Retrieved July 17, 2006, from http://news.bbc.co.uk/1/hi/technology/3906823.stm

BitDefender. (2004). *Proof-of-concept virus hits the last virus-resistant Microsoft OS.* Retrieved July 17, 2004, from http://www.bitdefender.com/bd/site/presscenter.php?menu_id=24&n_id=102

CIC. (1999). *Economic assessment office report: Computer recognition of natural handwriting.* Retrieved August 8, 2004, from http://statusreports-atp.nist.gov/reports/90-01-0210.htm

Cringely, R. X. (1996). *Accidental empires: How the boys of Silicon Valley make their millions, battle foreign competition and still can't get a date* (pp.164-167). New York: Penguin Books.

IBM. (2006). *ThinkPad: A brand that made history.* Retrieved August 8, 2006, from http://www.pc.ibm.com/us/thinkpad/anniversary/history.html

Karuturi, S. (2002). *SMS history.* Retrieved August 8, 2006, from http://www.funsms.net/sms_history.htm

Kay, A. (1972a, August). A personal computer for children of all ages. *Proceedings of the ACM National Conference* (pp. 370-376).

Kay, A. (1972b, November). A dynamic medium for creative thought. *Proceedings of the National Council of Teachers of English Conference* (pp. 121-124).

Kay, A., & Goldberg, A. (1977). Personal dynamic media. *IEEE Computer,* (March), 31-41.

Melnick, D., Dinman, M., & Muratov, A. (2004). *PDA security: Incorporating handhelds into the enterprise* (pp. 129-131). New York: McGraw-Hill.

Stanford University. (2003) *Human computer interaction: Designing technology.* Retrieved August 10, 2006, from http://hci.stanford.edu/cs547/abstracts/03-04/031003-moggridge.html

Troni, P., & Lowber, P. (2001). *Very portable devices (tablet and clamshell PDAs, smart phones and mini-notebooks: An overview.* Retrieved August 10, 2004, from http://cnscenter.future.co.kr/resource/rsc-center/gartner/portabledevices.pdf

KEY TERMS

Audio Memo: An audio recorded message of speech digitally recorded as an audio file on a PDA.

Laptop: A portable personal computer small enough to use on your lap.

Media Player: A device or software application designed to play a variety of digital communications media such as compressed audio files (e.g., MPEG MP3 files), digital video files, and other digital media formats.

Multimedia: Communications media that combines multiple formats such as text, graphics, sound, and video.

Multimedia Messaging Service (MMS): An emerging cellular phone service that allows the sending of multiple media in a single message, with the ability to send a message to multiple recipients. As such it can be seen as an evolution of SMS, with MMS supporting the transmission of additional media types, including: pictures, audio, video, and combinations of the above.

Palmtop: A portable personal computer which can be operated comfortably while held in one hand.

Pen Computing: A computer that uses an electronic pen (or stylus) rather than a keyboard for data input. Pen-based computers often support handwriting or voice recognition so that users can write on the screen or vocalize commands/dictate instead of typing with a keyboard. Many pen computers are handheld devices. Also known as pen-based computing.

Personal Digital Assistant (PDA): A small handheld computing device with data input and display facilities with a range of software applications. Small keyboards and pen-based input systems are commonly used for user input.

Personal Information Manager (PIM): A software application (such as Microsoft Outlook) that provides multiple ways to log and organize personal and business information such as contacts, events, tasks, appointments, and notes on a digital device.

Smart Phone: A term used for the combination of mobile phone and PDA.

Short Message Service (SMS): A text message service that enables users to send short messages (160 characters) to other users. A popular service amongst young people, with 400 billion SMS messages sent worldwide in 2002 (GSM World 2002).

Synchronization: The harmonization of data on two (or more) different digital devices so that both contain the same data. Data is commonly synchronized on the basis of the date it was last altered.

Tablet PC: A newer type of format for personal computers. The Tablet PC provides all the power of a laptop PC, but without a keyboard for text entry. Tablet PCs use pen-based input, and handwriting and voice recognition technologies as the main form of data entry, and commonly have an A4-size footprint.

Texting: Sending short text messages by SMS.

Wireless Connectivity: The communication of digital devices between one another using data transmission by radio waves.

Perceived Quality Evaluation for Multimedia Services

H. Koumaras
University of Athens, Greece

E. Pallis
Technological Educational Institute of Crete, Greece

G. Xilouris
University of Athens, Greece

A. Kourtis
N.C.S.R., Demokritos, Greece

D. Martakos
University of Athens, Greece

INTRODUCTION

The advent of 3G mobile communication networks has caused the fading of the classical boundaries between telecommunications, multimedia, and information technology sectors. The outcome of this convergence is the creation of a single platform that will allow ubiquitous access to the Internet, multimedia services, and interactive audiovisual services, and in addition (and most importantly) offering the required/appropriate perceived quality level at the end user's premises.

In this respect, multimedia services that distribute audiovisual content over 3G/4G mobile communication systems are expected to posses a major part of the bandwidth consumption, making necessary the use of video compression. Therefore, encoding techniques (e.g., MPEG, H-26x) will be applied which achieve high compression ratios by exploiting the redundancy in the spatiotemporal domain of the video content, but as a consequence produce image artifacts, which result in perceived quality degradation.

One of the 3G/4G visions is the provision of audiovisual content at various quality and price levels. There are many approaches to this issue, one being the perceived quality of service (PQoS) concept. The evaluation of the PQoS for audiovisual content will provide a user with a range of potential choices, covering the possibilities of low-, medium-, or high-quality levels. Moreover the PQoS evaluation gives the service provider and network operator the capability to minimize the storage and network resources by allocating only the resources that are sufficient to maintain a specific level of user satisfaction.

The evaluation of the PQoS is a matter of post-encoding procedures. The methods and techniques that have been proposed in the bibliography mainly aim at:

- determining the encoding settings (i.e., resolution, frame rate, bit rate) that are required in order to carry out successfully a communication task of a multimedia application (i.e., videoconference); and
- evaluating the quality level of a media clip based on the detection of artifacts on the signal caused by the encoding process.

The scope of this article is to outline the existing procedures and methods for estimating the PQoS level of a multimedia service.

BACKGROUND

The advent of quality evaluation was based on applying pure mathematical/error-sensitive equations between the encoding and the original/uncompressed video signal. These primitive methods, although they provided a quantitative approach about the quality degradation of the encoded signal, do not provide reliable measurements of the perceived quality, because they miss the characteristics and sensitivities of the human visual system.

The most widely used primitive methods and quality metrics that are based on the error sensitivity framework are the peak signal to noise ratio (PSNR) and the mean square error (MSE):

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

$\text{PSNR}=10\log_{10}\frac{L^2}{\text{MSE}}$, *where L denotes the dynamic pixel value (i.e., equal to 255 for 8bits/pixel monotonic signal)* (1)

$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2$, *where N denotes the total pixels and x_i /y_i the i^{th} pixel value in the original/distorted signal* (2)

Currently, the evaluation of the PQoS is a matter of objective and subjective evaluation procedures, each time taking place after the encoding process (post-encoding evaluation). Subjective picture/audio quality evaluation methods require a large amount of human resources, establishing it as a time-consuming process (e.g., large audiences evaluating video/audio sequences). Objective evaluation methods, on the other hand, can provide PQoS evaluation results faster, but require a large amount of machine resources and sophisticated apparatus configurations. Towards this, objective evaluation methods are based on and make use of multiple metrics, which are related to the content's artifacts (i.e., tilling, blurriness, error blocks, etc.) resulting during an encoding process.

These two categories of PQoS evaluation methods will be analyzed and discussed in the following sections.

SUBJECTIVE QUALITY EVALUATION METHODS

The subjective test methods, which have mainly been proposed by the International Telecommunications Union (ITU) and the Video Quality Experts Group (VQEG), involve an audience of people who watch a video sequence and score its quality, as perceived by them, under specific and controlled watching conditions. Afterwards, the statistical analysis of the collected data is used for the evaluation of the perceived quality. The mean opinion score (MOS) is regarded as the most reliable subjective metric of quality measurement and has been applied on the most known subjective techniques.

Subjective test methods are described in ITU-R Rec. T.500-11 (2002) and ITU-T Rec. P.910 (1999), suggesting specific viewing conditions, criteria for the observer, test material selection, assessment procedure description, and statistical analysis methods. The BT.500-11 describes subjective methods that are specialized for television applications, whereas ITU-T Rec. P.910 is intended for multimedia applications.

The most known and most widely used subjective methods are:

- **Double Stimulus Impairment Scale (DSIS):** This method proposes that observers watch multiple references and degraded scene pairs, with the reference scene always shown first. Scoring is evaluated on an overall impression scale of impairment: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying. This scale is commonly known as the five-point scale (where 5 corresponds to "imperceptible" and 1 to "very annoying").
- **Single Stimulus (SS) Methods:** Multiple separate scenes are shown. There are two different SS approaches: SS with single view of test scenes and SS where the test scenes are repeated. Three different scoring methods are used:
    - **Adjectival:** The aforementioned five-grade impairment scale, however half-grades are allowed.
    - **Numerical:** An 11-grade numerical scale, useful if a reference is not available.
    - **Non-Categorical:** A continuous scale with no numbers or a large range, for example, 0-100.
- **Stimulus Comparison Method:** This methods exploits two well-matched screens, where the differences between scene pairs are scored in one of the two following scoring methods:
    - **Adjectival:** A seven-grade, +3 to -3 scale labeled: much better, better, slightly better, the same, slightly worse, worse, and much worse.
    - **Non-Categorical:** A continuous scale with no numbers or a relation number either in absolute terms or related to a standard pair.
- **Single Stimulus Continuous Quality Evaluation (SSCQE):** According to this method, the viewers watch a program of typically 20-30 minutes without any reference signal. The viewers, using a slider, continuously rate the instantaneously perceived quality using an adjectival scale from 'bad' to 'excellent', which corresponds to an equivalent numerical scale from 0 to 100.
- **Double Stimulus Continuous Quality Scale (DSCQS):** At DSCQS the viewers watch multiple pairs of quite short (i.e., 10 seconds) reference and test sequences. Each pair appears twice, with random order of the reference and the test sequence. The viewers/subjects are not aware of the reference/test order, and they are asked to rate each of the two separately on a continuous adjectival scale, ranging from 'bad' to 'excellent', which corresponds to an equivalent numerical scale from 0 to 100. This method is usually used for evaluating slight quality differences between the test and the reference sequence.

The aforementioned methods are described in the ITU-R Rec. T.500-11 document and are mainly intended for television signals. Based on slight modifications and adaptations of these methods, some other subjective evaluation methods (namely absolute category rating (ACR), degradation category rating (DCR), etc.) for multimedia services are described in ITU-T Rec. P.910.

OBJECTIVE QUALITY EVALUATION METHODS

The preparation and execution of subjective tests is costly and time consuming, and its implementation today is limited to scientific purposes, especially at VQEG experiments.

For this reason, a lot of effort has recently been focused on developing cheaper, faster, and more easily applicable objective evaluation methods. These techniques successfully emulate the subjective quality assessment results, based on criteria and metrics that can be measured objectively. The objective methods are classified according to the availability of the original video signal, which is considered to be of high quality.

The majority of the proposed objective methods in the literature require the undistorted source video sequence as a reference entity in the quality evaluation process, and due to this are characterized as full reference methods. These methods emulate characteristics of the human visual system (HVS) using contrast sensitivity functions (CSF), channel decomposition, error normalization, weighting, and finally Minkowski error pooling for combining the error measurements into single perceived quality estimation (Wang, Sheikh, & Bovik, 2003).

However it has been reported (VQEG, 2000; Wang, Bovik, & Lu 2002) that these complicated methods do not provide more accurate results than the simple mathematical measures (such as PSNR). Due to this some new full reference metrics that are based on the video structural distortion, and not on error measurement, have been proposed (Wang et al., 2003).

On the other hand, the fact that these methods require the original video signal as reference deprives their use in commercial video service applications, where the initial undistorted clips are not accessible. Moreover, even if the reference clip is available, then synchronization predicaments between the undistorted and the distorted signal (which may have experienced frame loss) make the implementation of the full reference methods difficult and impractical.

Due to these reasons, recent research has focused on developing methods that can evaluate the PQoS level based on metrics, which use only some extracted structural features from the original signal (Reduced Reference Methods—Guawan & Ghanbari, 2003) or do not require any reference video signal (No Reference Methods—Lu, Wang, Bovik, & Kouloheris, 2002).

However, due to the fact that the 3G/4G vision is the provision of audiovisual content at various quality and price levels (Seeling, Reisslein, & Kulapala, 2004), there is great need for developing methods and tools that will help service providers to predict quickly and easily the PQoS level of a media clip. These methods will enable the determination of the specific encoding parameters that will satisfy a certain quality level. All the previously mentioned post-encoding methods may require repeating tests in order to determine the encoding parameters that satisfy a specific level of user satisfaction. This procedure is time consuming, complex, and impractical for implementation on the 3G/4G multimedia mobile applications.

Towards this, recently research was performed in the field of pre-encoding estimation and prediction of the PQoS level of a multimedia service as a function of the selected resolution and the encoding bit rate (Koumaras, Kourtis, & Martakos, 2005; Koumaras et al., 2004). These methods provide fast and quantified estimation of the PQoS, taking into account the instant PQoS variation due to the spatial and temporal (S-T) activity within a given encoded sequence. Quantifying this variation by the mean PQoS (MPQoS) as a function of the video encoding rate and the picture resolution, it finally used the MPQoS as a metric for pre-encoding PQoS assessment based on the fast estimation of the S-T activity level of a video signal.

FUTURE TRENDS

Simultaneously with the development of the aforementioned methods and techniques, research has been focused on developing methods that determine the adequate quality level for a specific multimedia application, taking under consideration not solely the visual estimations, but also a great number of parameters and metrics that depend on the task nature and the user emotional behavior and psychophysical characteristics (Mullin, Smallwood, Watson, & Wilson, 2001). For example, the classification of the task as foreground or background in correlation with its complexity (Buxton, 1995) is a parameter that differentiates the quality demands of a multimedia application. On the other hand, the emotional content of a multimedia communication task alters the required quality level of the specific communication service (Olson, 1994). Due to this, various parameters are measured in order to estimate the appropriate minimum quality level of a multimedia application. Such parameters are:

- the user characteristics (i.e., knowledge background, language background, familiarity with the task, age);
- the situation characteristics (i.e., geographical remoteness, simultaneous number of users, distribution of users);

- the user cost (i.e., heart rate, blood volume pulse); and
- the user behavior (i.e., eye tracking, head movement).

However, these methods still have some issues to solve on the technical, theoretical, and practical levels. A user that participates in such an assessment procedure is so wired (even on the head, he or she may wear the eye tracking equipment) that it causes uncomfortable feelings and affects his or natural behavior. Technical issues, such as the eye tracking loss and the manual calibration/correction by a human operator, affect the reliability of the methods in real-time environments (Mullin et al., 2001).

CONCLUSION

Multimedia applications, and especially encoded video services, are expected to play a major role in third-generation (3G) and beyond mobile communication systems. Given that future service providers are expected to provide video applications at various price and quality levels, quick and economically affordable methods for preparing/encoding the offering media at various qualities need to be developed. There are a number of approaches to this challenge, one being the use of the perceived quality of service concept. The evaluation of the PQoS for multimedia and audiovisual content that has variable bandwidth demands will provide a user with a range of choices covering the possibilities of low-, medium-, or high-quality connections, an indication of service availability and cost. This article outlines the various existing PQoS evaluation methods and comments on their efficiency.

These methods can be mainly categorized into two major classes: subjective and objective. The subjective test methods involve an audience of people who watch a video sequence and evaluate its quality, as perceived by them, under specific and controlled watching conditions. The objective methods successfully emulate the subjective quality assessment results, based on criteria and metrics that can be measured objectively. These objective methods are classified according to the availability of the original video signal to full reference, reduced reference, and no reference.

However, all the aforementioned post-encoding methods require repeating post-encoding tests in order to determine the encoding parameters that satisfy a specific level of user satisfaction, making them time consuming, complex, and impractical for implementation on the 3G/4G multimedia mobile applications. Due to this, lately some new pre-encoding evaluation methods have been proposed that are capable of estimating/predicting the PQoS level of a multimedia service based on the selected resolution, bit rate, and content activity. These methods quickly provide accurate estimations of PQoS level, alleviating the time and resource requirements that the traditional objective methods consume.

ACKNOWLEDGMENTS

The work in this article was carried out in the frame of the Information Society Technologies (IST) project ENTHRONE/FP6-507637.

REFERENCES

Buxton, W. (1995) Integrating the periphery and context: A new taxonomy of telematics. *Proceedings of Graphics Interface 1995* (pp. 239-246).

Guawan, I. P., & Ghanbari, M. (2003). Reduced-reference picture quality estimation by using local harmonic amplitude information. *Proceedings of the London Communications Symposium 2003.*

Koumaras, H., Kourtis, A., & Martakos, D. (2005). Evaluation of video quality based on objectively estimated metric. *Journal of Communications and Networking, 7*(3), 235-242.

Koumaras, H., Pallis, E., Xilouris, G., Kourtis, A., Martakos, D., & Lauterjung, J. (2004). Pre-encoding PQoS assessment method for optimized resource utilization. *Proceedings of the 2nd International Conference on Performance Modeling and Evaluation of Heterogeneous Networks (Het-NeTs04)*, Ilkley, UK.

Lu, L., Wang, Z., Bovik, A. C., & Kouloheris, J. (2002). Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video. *Proceedings of the IEEE International Conference on Multimedia.*

Mullin, J., Smallwood, L., Watson, A., & Wilson, G. (2001). New techniques for assessing audio and video quality in real-time interactive communications. *Proceedings of the 3rd International Workshop on Human Computer Interaction with Mobile Devices,* Lille, France.

Olson, J. (1994). In a framework about task-technology fit, what are the tasks features? *Proceedings of CSCW '94: Workshop on Video Mediated Communication: Testing, Evaluation & Design Implications.*

Seeling, P., Reisslein, M., & Kulapala, B. (2004). Network performance evaluation using frame size and quality traces of single layer and two layer video: A tutorial. *IEEE Communications Surveys, 6*(3).

VQEG. (2000). *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment.* Retrieved from http://www.vqeg.org

Wang, Z., Bovik, A. C., & Lu, L. (2002). Why is image quality assessment so difficult? *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing* (Vol. 4, pp. 3313-3316).

Wang, Z., Sheikh, H. R., & Bovik, A. C. (2003). Objective video quality assessment. In B. Furht & O. Marqure (Eds.), *The handbook of video databases: Design and applications* (pp. 1041-1078). CRC Press.

KEY TERMS

Bit Rate: A data rate expressed in bits per second. In video encoding the bit rate can be *constant,* which means that it retains a specific value for the whole encoding process, or *variable,* which means that it fluctuates around a specific value according to the content of the video signal.

Double Stimulus Continuous Quality Scale (DSCQS): A subjective evaluation method according to which the viewers watch multiple pairs of quite short (i.e., 10 seconds) reference and test sequences. Each pair appears twice, with random order of the reference and the test sequence.

Multimedia: The several different media types (e.g., text, audio, graphics, animation, video).

Objective Measurement of Perceived Quality: A category of assessment methods that evaluates the PQoS level based on metrics, which can be measured objectively.

Perceived Quality of Service (PQoS): The perceived quality level that a user experiences from a multimedia service.

Quality Degradation: The drop of the perceived quality to a lower level.

Single Stimulus Continuous Quality Evaluation (SSCQE): A subjective evaluation method according to which the viewers watch a program of typically 20-30 minutes, without the original reference shown, and score its quality.

Spatial-Temporal Activity Level: The dynamics of the video content, in respect to its spatial and temporal characteristics.

Pest Activity Prognosis in the Rice Field

Nureize Arbaiy
Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

Azizul Azhar Ramli
Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

Zurinah Suradi
Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

Mustafa Mat Deris
Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

INTRODUCTION

In crops management, it is important to estimate the damage effected by pests since the degree of damage will determine the level of pest activity. Pest activity usually involves their life stage and its presence in the field. In addition, pest management in crops is a crucial problem and may yield losses if it is not handled properly. Consequently a forecasting tool is needed to predict the level of pest activity. This is important so that an early treatment or action can be applied before more damage to the plant occurs.

Accordingly, the fuzzy expert system may facilitate the user through a consultation session in order to forecast the pest activity in the rice field. A set of questions will be asked to help users diagnose their given symptom in order to infer such a conclusion. Figure 1 shows the main components of an expert system including inference engine, expert, knowledge base, working memory, and user interface. The consultation performed by the expert system also involves fuzzy logic to deal with the natural and uncertainty data. Besides, all the information and knowledge about the pests, treatment control measures and prevention steps are managed in the specific knowledge base created in the system. This system is able to educate and inform the farmers and smallholders about pests and their activities in the rice field.

BACKGROUND

Sustainable agriculture is a key element of sustainable development and essential to the future well-being of the planet. Sustainability aims to achieve adequate safe and healthy food production, improved livelihoods of food producers, and the preservation of non-renewable resources. The demands of a growing world population for food and fiber require world agriculture to produce higher yields from less cultivated land. Recently, the emerging need for hybrid intelligent systems (HISs) is motivating important research and development work. The integration of different learning and adaptation techniques to overcome individual limitations and achieve synergetic effects through hybridization or fusion of these techniques has in recent years contributed to a large number of new intelligent system designs (Abraham & Nath, 2001). This integration aims at overcoming limitations of individual techniques through hybridization of fusion of various techniques.

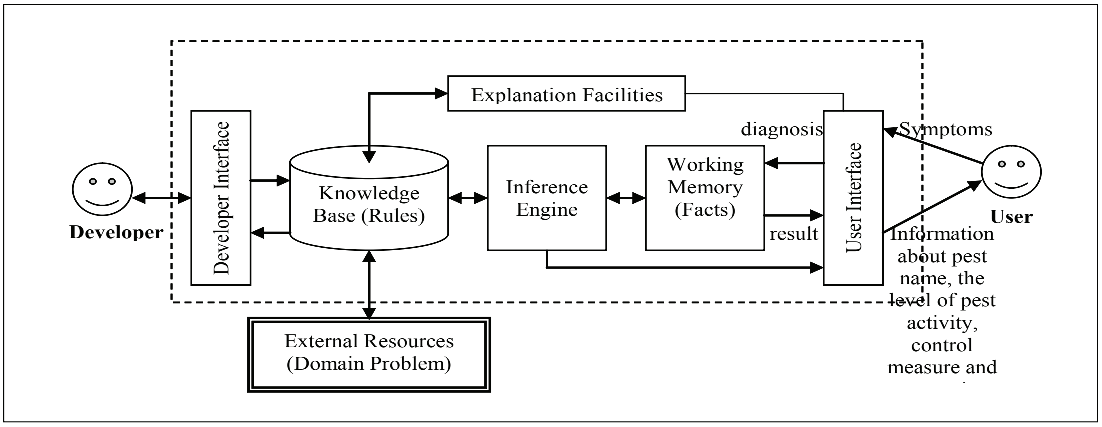
In the past decade, a great many expert systems—such as office automation, science, and medicine including agriculture—were developed and applied to many fields. At the initial stage of agriculture expert system development, the focus was on diseases diagnosis and pests of various crops. In recent years, research and development of the expert system fields of the agriculture domain received much attention by many countries. The difficulty of problems confronting farmers are yield losses, soil erosion, diminishing market prices from international competition, increasing chemical pesticide costs, pest resistance, and economic barriers. Hence, farmers need such an application or tools to assist with various knowledge and information, especially in their farming operations.

PEST ACTIVITY PROGNOSIS

Pest management in agriculture is a pest control strategy that coordinates and uses a combination of methods to meet individuals' production goals in the most economical and environmental manner. In short, it is based on understanding the farm's ecology. The basic components of pest management include acceptable pest levels, preventive cultural practices, monitoring, mechanical controls, biological control, and chemical control. It is apparent that that farm's monitoring and prevention endow with significant contribution to reduce losses.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Expert system components



The management of pests in crops is a highly challenging problem and may yield losses if not handled properly (Saini, Kamal, & Sharma, 2002). Potential losses of up to 55% before harvest have been estimated, but these estimates often represent the worst case or highest levels of loss. Consequently, different technologies are needed as well as awareness programs for effective, economical, environment-friendly control of pests (Singh & Singh, 1990). Besides, the appropriate and optimal combination of control measures are used for cost-effective and environment-friendly control of pests (Atwal & Dhaliwal, 1997).

It is important to manage and control the pest occurrences and their activity in the crop fields. Pests in crop management involves birds, snails, worms, rats, and others. Good field practices to protect and sustain the field require proficient knowledge and information. Here, farmers are encouraged to learn their farm, observe and monitor, and use their knowledge and experience to decide on actions. They may also seek advice from farm experts (researchers and governmental representatives). Consequently, the transfer of expert knowledge and extension information to the farmers plays a key role in educating this community. The understanding of the farmers is further increased through experimentation and knowledge sharing. For that reason, an expert system seems to be beneficial to vary the knowledge sharing and reasoning. Expert system offer a program that uses available information, heuristics, and inference to suggest solutions to problems in a particular discipline. While such systems do not often replace the human experts, they can serve as useful assistants. Expert systems will play a major role in the dissemination and application of useful knowledge leading to economic growth and higher standards of living. It is not only providing expert knowledge, but potentially become learning resources to help farmers develop their own expertise.

On the other hand, as knowledge involved in pest management is imperfect and fuzzy logic has been successfully used for approximate reasoning in such cases, its application becomes mandatory to manage the uncertainty in the expert system (Zadeh, 1983). Appearance of damage at a farm does not always promise identification of the pests. Crop damage is never complete and usually expressed linguistically as very low, low, medium, high, and very high. Moreover, the partial truth values of existing pest symptoms may strengthen pest identification. The value of pest occurrence symptoms is also captured in fuzzy terms such as few, many, light, and many more. Due to the imperfect, vague, and not completely reliable knowledge involved in pest activity and damage level in the rice fields, it is difficult to measure the symptom occurrences with simply yes or no, or absence and presence notation. The crisp rules may not be precisely appropriate for pest identification (Pasqual & Mansfield, 1988). For instance, the existence of larvae in the leaves cannot be simply expressed by the 'yes' or 'no' value.

However, the existing expert system allows the user to answer the set of questions using the rigid crisp values (Saini et al., 2002). In crops management, it is important to estimate the damage that has been affected by pests, since the degree of damage will determine the activity of pests (Atwal & Dhaliwal, 1997). Therefore, there is need for a forecasting tool that can predict the level of pest activity so that early treatments can be applied to crops before the damage becomes worse. Hence fuzzy logic helps to cope with the precise damage symptoms for pest activity estimation, which is proportional to the level of damages. Fuzzy modeling techniques such as fuzzy sets, fuzzy logic, and fuzzy expert systems are used to capture vague concepts and process imprecise information, and can further implement and express human knowledge and inference capability in a natural way. The design of a fuzzy system mainly involved two operations: the derivation of the knowledge base, and the selection of the fuzzy inference with a defuzzification process that the system will use to perform the fuzzy reasoning (Cordon, Herera, & Peregrin, 1994).

In order to develop such an intelligent system embedded with some prediction and forecasting ability, a lot of effort is needed. The advantage of fuzziness is that dealing with imprecision fits ideally into decision systems. The vagueness and uncertainty of human expressions is well modeled in the fuzzy sets, and a pseudo-verbal representation, similar to an expert's formulation, can be achieved (Hasiloglu, Yavuz, Rezos, & Kaya, 2003). Garibaldi (1997) stated that in multi-valued logic, truth values are represented by a single real number in the interval [0,1], where 0 represents false, 1 represents true, and values between 0 and 1 represent partial truth, whereas in fuzzy logic, true and false are represented by fuzzy subsets over the interval [0.1], with arbitrary fuzzy subsets representing other intermediate truth values.

Fuzzy Expert System

Hybrid systems composed of artificial intelligence (AI) approaches have shown quite remarkable results in diagnosis (Herrmann, 1995). A fuzzy expert system is an expert system that uses fuzzy logic instead of Boolean logic, and a collection of membership functions and rules that are used to reason about data. Unlike conventional expert systems, fuzzy expert systems are oriented toward numerical processing. This expert system is extended to incorporate explicit handling of imprecision in the input data and uncertainty in the embedded knowledge. The system formed the basis of the fuzzy expert system. The existent crisp rule set was used to derive the initial fuzzy rule set and to guide the initial choice of location of membership function for each fuzzy term (Garibaldi et al., 1999).

In contrast with conventional AI techniques, which only deal with precision and uncertainty, the guiding principle of hybrid systems is to exploit the tolerance for imprecision, uncertainty, low solution cost, robustness, partial truth to achieve tractability, and better understanding with reality (Zadeh, 1998). However, although the theoretical properties of fuzzy systems have been extensively investigated, the implementation of a fuzzy expert system in practice involves a great deal of pragmatic choices. This includes considerations for the type of inference methodology, rule set, and fuzzy operators to determine an appropriate fuzzy model of the expertise for a particular application.

Moreover, the prognosis term has been used widely in medical applications and defined as a forecast of course of disease (Coulson, Carr, Hutchinson, & Eagle, 1990). Since prognosis requires forecasting ability as well as the ability to explain why a phenomenon occurs, AI techniques that are required to perform prognosis must be able to forecast and provide reasoning. AI techniques that are suitable for prediction are neural network, fuzzy logic, and case-based reasoning. On the other hand, expert system and case-based reasoning are good at providing explanation to intelligent system. In this study, techniques which have the forecasting and explanation abilities are required. For this purpose, fuzzy logic and an expert system have been chosen to be integrated into a Web-based environment to demonstrate the used of a hybrid system on pest activity in rice field data.

MOBILE COMPUTING AND CELLULAR TECHNOLOGY

Wireless technologies represent a rapidly emerging area of growth and importance for providing ubiquitous access to the network for various communities. The most related network technology with wireless technology is the mobile computing environment. Mobile computing systems are computing systems that may be easily moved physically and may be used while users are on the move. Examples of mobile devices are laptops, personal digital assistants (PDAs), and mobile phones.

The initiation of portable computers and handheld devices simultaneously with Internet technology has led to mobile computing. The growth of mobile devices has tremendously increased its usability and provides more easy access. As devices become smaller and more portable, the demand for computing and networking solutions while on the move has increased steadily. In the mobile computing environment, users can perform online transaction processing independent of the physical location. Geographical constraints can also be eliminated from data processing activities. Accordingly, data and information stored in a certain database can be accessed widely, using mobile technology and devices. This is important to disseminate information and knowledge among the users.

However, mobile computing environments must deal with limited and dynamically varying resources, in particular, the network quality of service (Bharghavan & Gupta, 1997). This shows that, to facilitate the success of mobile computing application, many factors should be considered. These include slow networks, wasteful protocols, disconnections, weak terminals, operating systems (OSs), and many others (Satyanarayanan, 1995). Nevertheless many efforts have progressed to improve the constraints and increase the usability. For example, enhancements in applications and devices and increasing of data rates will drive to the implementation of the most efficient solution such as 3G. New devices will be introduced and existing devices will evoke to provide more enhanced wireless data experience.

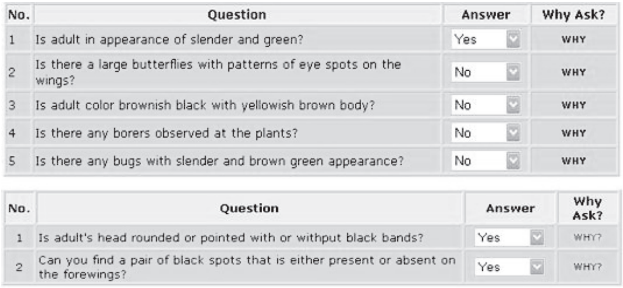
Mobility applications enable the data transmission and user interaction between wireless devices and a central repository. Data transmission is handled by synchronization applications in situations where the user is not connected to the central database on a real-time basis (Ramadhani & Siddiqi, 2003). Facilitated with numerous handheld devices, the data transmission through these devices extend the mobile application, especially in distributing information among citizens.

Basically, a cellular network is a radio network made up of a number of radio cells, each served by a fixed transmitter known as a cell site or base station. These cells are used to cover different areas in order to provide radio coverage over a wider area than the area of one cell. The most common example of a cellular network is a mobile phone (cell phone) network. A mobile phone is a portable telephone that receives or makes calls through a cell site (base station) or transmitting tower. Mobile phones are becoming more powerful and are equipped with additional large memory, networking interfaces (i.e., IrDA, Bluetooth, and GSM/GPRS), and can reconcile packets between the different networks. Hence, to being able to connect to the network operator, modern phones must have network connection facility. Different hardware manufacturers usually provide different device features (i.e., operating and middleware systems) to the users. Therefore, such particular features are important to support the deployment of the system application using a mobile phone's capability.

FUZZY EXPERT SYSTEM APPLICATION USING MOBILE PHONE

This study focuses on the software development using hybrid AI technology and the employment of a fuzzy expert system in the agriculture domain typically in Malaysia. Therefore it emphasizes data acquisition and mapping of uncertainty into fuzzy values, which consists of labels and confidence values. The application offers a computerized fuzzy expert system in dealing with uncertainty information in a way to identify the kinds of pests that attack the rice plant derived from the symptoms given by the farmers. The system helps the user by managing the consultation session in order to forecast the pest activity. With reference to Figure 2, the user has given answers to the related questions and the system will invoke the knowledge base to find the conclusion based on the given answer.

Figure 2. Set of expert questions



No.	Question	Answer	Why Ask?
1	Is adult in appearance of slender and green?	Yes	WHY
2	Is there a large butterflies with patterns of eye spots on the wings?	No	WHY
3	Is adult color brownish black with yellowish brown body?	No	WHY
4	Is there any borers observed at the plants?	No	WHY
5	Is there any bugs with slender and brown green appearance?	No	WHY

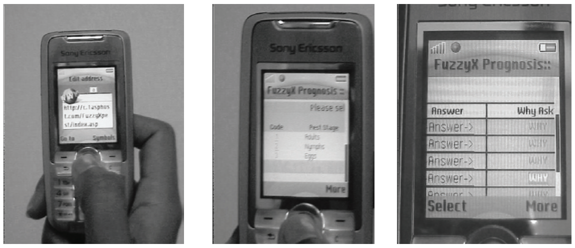
No.	Question	Answer	Why Ask?
1	Is adult's head rounded or pointed with or withput black bands?	Yes	WHY?
2	Can you find a pair of black spots that is either present or absent on the forewings?	Yes	WHY?

This will help the user diagnose his or her given symptom in order to infer such a conclusion. The consultation performed by the expert system also involved fuzzy logic when dealing with the natural and uncertainty data. The occurrences of pests in the rice field and crop production may affect the field yields. Pest management is needed where losses are much higher, resulting in complete damage. These pests attack the crop from the time of sowing through to harvest and beyond, often causing significant economic losses. Therefore a forecasting system is intended to facilitate the warning system so farmers can apply early treatment to avoid high losses resulting from complete damage.

Even though such a prognostic system has been developed, it is important to distribute and make the application accessible by these farmers. Consequently, this system will facilitate as well as educate the farmers to possess their awareness in their own field management. Moreover, this prognostic application can be accessed using an affordable mobile device. The mobile devices equipped with minimal 2.5G application (GPRS connection) will support the prognostic application. Figure 3 shows the access of prognostic application through mobile phones.

Thus, the prognostic system should be made more accessible to the farmers using mobile devices. Mobile users are able to access and exchange information while they are on the move. Meanwhile, mobile devices need a mobile computing technology to wirelessly connect and use centrally located information and application software. Mobile phones, for instance, are easy to use and equipped with minimal setting of wireless network configuration. Availability and network coverage in these devices depend upon the carrier and the

Figure 3. Accessing from mobile phone



geographic scope of international roaming agreements. Therefore it can allow for a more effective, convenient, and timely use of computing and communication. It is expected that the wireless technology will play an important role in the acquisition and dissemination of new knowledge and technologies to motivate the involvement of youth in the agricultural sector. It is also anticipated to redress the digital divide between the farming community and others.

FUTURE DIRECTIONS

This study attempts to explore the employment of a fuzzy expert system in the agriculture domain typically in Malaysia. Previous study performed by Saini et al. (2002) has been developed to provide pest management to the farmers through the Internet. Saini et al. (2002) have also improved the existing expert system they developed by applying fuzzy logic to handle the uncertainty involved in damage forecasting. Since the system produced is a prototype, the system needs more enhancement and improvements in order to get better results and performance. For the forecasting tasks, if more historical data can be captured, it is beneficial to integrate neural network and fuzzy logic. Neuro-fuzzy systems are known to mitigate the limitations and take advantage of opportunities to produce more powerful hybrids than those that could be built with stand-alone systems where it facilitates fast learning and online adaptability, and achieves a global error rate and computational inexpensive.

On the other hand, an intelligent agent approach can be applied to support Web facilities as well as provide further explanation for expert system. It is a possible way to increase explanation facilities by integrating an intelligent agent into this system, perhaps using case-based reasoning. This indirectly will enhance the support desk facilities for the users. Further enhancement can also be expanded to the other crop management systems. The same engine still can be used, however the database for other crops needs to be developed. In other words, community users will be able to benefit from the system.

CONCLUSION

This study focuses on the employment of a fuzzy expert system in order to make a prognosis on the pest level of activity in the rice fields. Since the rice industry has been and is likely to continue to be one of the most important economic sectors of the company, there is a need to help enhance the competitiveness and profitability in agriculture and forestry in this country. The crop is labor intensive and needs some portion of expenditure to manage this sector. Recently, most of the farmers in this region are gradually moving away from the image of the traditional farmer to that of an entrepreneur farmer. This success technology can be implemented in some areas to gain benefits from that.

Furthermore, the prognostic application can be accessed using a mobile phone. This is important to disseminate agriculture's knowledge among farmers through mobile devices. This will also facilitate farmers with such an assisting prognostic tool to help manage their fields and to develop awareness of field monitoring from threats and disease. The prognostic system's application allows users to input a percentage of symptoms in uncertainty forms (high, very high, medium) rather than the common form of yes/no or absence/presence form. The system enables the users, particularly the farmers and the governmental representatives, to identify the pest that damages the plant. The system also allows the users to forecast a pest activity level in the rice field and provide the treatment information before the pest activities become worse. In addition, all the information and knowledge about the pests, treatment control measures, and prevention steps are managed in the specific knowledge base. Apart form identifying the pest and its activity level, the system can be used as part of portal development for agricultural agencies in particular and the farmer community as a whole.

REFERENCES

Abraham, A., & Nath, B. (2001). *Hybrid intelligent systems design—A review of a decade of research.*

Agrawal, P., & Famolari, D. (1999). *Mobile computing in next generation wireless networks.* Swedish Institute of Computer Science, Sweden.

Atwal, A. S., & Dhaliwal, G. S. (1997). *Agricultural pests of south Asia and their management.* Kalyani Publisher.

B'Far, R. (2001). Mobile computing principles: Designing and developing mobile applications with UML and XML. Cambridge: Cambridge University Press.

Bharghavan, V. & Gupta, V. (1997). A framework for application adaptation in mobile computing environments. *Proceedings of IEEE Compsac 1997* (pp. 573-579).

Cordon, O., Herera, F., & Peregrin, A. (1999). *Looking for the best defuzzification method features for each implication operator to design accurate fuzzy model.* University of Granada, (technical report DECSAI-99108). Department of Computer Science and Artificial Intelligent, Spain.

Coulson, J., Carr, C. T., Hutchinson, L., & Eagle, D. (1990). *The English illustrated dictionary.* Oxford: Oxford University.

Garibaldi, J. M. (1997). *Intelligent techniques for handling uncertainty in the assessment of neonatal outcome.* PhD thesis, University of Plymouth, UK.

Hadjimichael, M., Kuciauskas, A. P., Brody, L. R., Bankert, R. L., & Tag, P. M. (1996). MEDEX: A fuzzy system for forecasting Mediterranean gale force winds. *Proceedings of the FUZZ-IEEE 1996 IEEE International Conference on Fuzzy Systems* (pp. 529-534).

Hasiloglu, A. S., Yavuz, U., Rezos, S., & Kaya, M. D. (2003). A fuzzy expert system for product life cycle management. In *Proceedings of the 12th International Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2003)*.

Herrmann, C.S. (1995). *Fuzzy logic as interfacing technique in hybrid AI-systems.* Germany.

Kuciauskas, A. P., Brody, L. R., Hadjimichael, M., Bankert, R. L., & Tag, P. M. (1998). MEDEX: Applying fuzzy logic to a meteorological expert system. *Proceedings of the 1st Conference on Artificial Intelligence* (pp. 68-74). American Meteorological Society.

Maner, W., & Joyce, S. (1997). WXSYS: Weather lore + fuzzy logic = weather forecasts. *Proceedings of the 1997 CLIPS Virtual Conference.* Retrieved February 27, 1999, from http://web.cs.bgsu.edu/maner/wxsys/wxsys.htm

Murtha, J. (1995). Applications of fuzzy logic in operational meteorology. *Scientific Services and Professional Development Newsletter,* 42-54. Canadian Forces Weather Service.

Pasqual, G. M., & Mansfield, J. (1998). Development of a prototype expert system for identification and control of insect pests. *Computer and Electronics in Agriculture, 2,* 263-276.

Radhamani, G., & Siddiqi, M. U. (2003). *An efficient WAP-enabled transaction processing model for mobile database systems.* CRC Press.

Saini, H. S., Kamal, R., & Sharma, A. N. (1997). Graphical user interface for a fuzzy expert system SOYPEST. *Vivek: A Quarterly in Artificial Intelligence, 10*(4), 2-10.

Saini, H. S., Kamal, R., & Sharma, A. N. (2002). Web-based fuzzy expert system for integrated pest management in soybean. *International Journal of Information Technology, 8*(1).

Satyanarayanan, M. (1995). *Fundamental challenges in mobile computing.* School of Computer Science, Carnegie Mellon University, USA.

Singh, O. P., & Singh, K. J. (1990). Insect pests of soybean and their management. *Indian Farming,* (January), 9-38.

Sujitjorn, S., Sookjaras, P., & Wainikorn, W. (1994, October 2-5). An expert system to forecast visibility in Don-Muang Air Force Base. *Proceedings of the 1994 IEEE International Conference on Systems, Man and Cybernetics (Humans, Information and Technology)* (pp. 2528-2531). New York.

Visser, M.A., & El Zarki, M. (1995). Voice and data transmission over an 802.11 wireless network. *Proceedings of PIMRC'95* (pp. 648-652), Toronto, Canada.

Zadeh, L. A. (1983). The role of fuzzy logic in management of uncertainty in expert systems. *Fuzzy Sets Systems, 11,* 199-227.

Zadeh, L. A. (1998). Roles of soft computing and fuzzy logic in the conception, design and deployment of information/intelligent systems. In O. Kaynak et al. (Eds.), *Computational intelligence: Soft computing and fuzzy neuro integration with applications* (pp. 1-9). Springer Verlag, Germany.

Positioning Technologies for Mobile Computing

Michael J. O'Grady
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

INTRODUCTION

Mobility is, as the name suggests, the defining characteristic of mobile computing and the primary differentiator between it and other computer usage paradigms. Traditionally, computers were used in what may be termed a static context. However, when computers are used in a mobile context, a number of difficulties that challenge traditional assumptions emerge. Not least amongst these are those difficulties that arise in delivering a service that is relevant and consistent with the situation in which the end-user find themselves. Should a person be waiting at a bus stop, he or she does not wish to go online and browse a bus timetable. Rather, he or she wishes to know when the next bus will stop at his or her particular stop. Thus location and time would be fundamental to the provision of such a service. Capturing time provides no major difficulties. However, identifying the physical location of a service subscriber may prove problematic.

In this review, we summarize some of the key technologies that enable the position of a mobile computer user to be determined.

BACKGROUND

Research in mobile computing and associated disciplines (Vasilakos, 2006) began in earnest the 1990s as the feasibility of the paradigm became increasingly clear. As the various research issues began to crystallize, researchers became aware of the desirability of using additional known facts of the end user's prevailing circumstances as a basis for customizing or personalizing the service for the individual end user. The term *context-aware computing* was coined to conceptualize these ideas. Pioneering research in this area was conducted at Xerox Parc in California by Schilit Adams, and Want (1996). The Oxford Concise dictionary defines context as "the interrelated conditions in which something exists or occurs." Intuitively, everybody understands what context is. Almost paradoxically, this has made the derivation of an agreed definition almost impossible, leading some researchers to reconsider its philosophical roots (Dourish, 2004) and inherently dynamic nature (Greenberg, 2001). One issue commonly agreed is that a person's location or physical position forms an indispensable aspect of his or her context—so much so that Schmidt, Beigl, and Gellerson (1999) almost remind researchers that there are other aspects of context that should be considered. The reasons for researchers' enthusiasm are understandable. In the mid-1990s, the global positioning system (GPS) was deployed, making it possible to determine position to within 100 meters for those people equipped with a GPS receiver. Thus the technological issues were being addressed in a meaningful way. However, it was developments in wireless telecommunications that provided the spur for the upsurge in business interest in what would be termed location-aware computing (Patterson, Muntz, & Pancake, 2003).

In 1996, the Federal Communications Commission (FCC) in the United States announced the E-911 directive. In brief: this obliged public telecommunication network operators to provide the position of those people making emergency calls, thus enabling police, medical, and other personnel to react quicker. It soon became clear that this facility could have other uses for commercial purposes as, in principle at least, the location of any subscriber could be identified. Thus an era of location-aware services was anticipated. This era has yet to materialize, but as outstanding technological issues are continually being addressed, it is only a matter of time before a suite of location-aware services are available for subscribers.

To deliver location-aware services, it is necessary that an appropriate technology be selected that will provide a subscriber's position within a certain range. In the next section, some of the principal technologies for determining position are described.

TECHNOLOGIES

Various technologies and techniques are described in the academic literature for determining user position. Naturally, each has its respective advantages and disadvantages. For the purposes of this discussion, it is useful to classify them as

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

satellite techniques, cellular network techniques, and hybrid. Each classification is now considered briefly.

Satellites Technologies

Trilateration is the basic principle for determining position using satellites. In short, the time taken for a signal to travel from a satellite at a known position to a receiver is calculated. This process is repeated for three satellites and a solution can be generated. In practice, a fourth measurement is necessary to account for the lack of synchronization between the atomic clocks on the satellite and the receiver's internal clock. The accuracy of the resultant calculation may vary due to a number of factors, including atmospheric conditions and the satellite constellation configuration. However, a reading within 20 meters of the receiver's exact geographic position may be realistically expected.

At present, there are two satellite systems in operation that broadcast signals:

1. *Global positioning system* was deployed in 1996, covers the entire earth, and is freely available. It remains under the control of the United States military. It is currently the de facto standard with specialized receivers on the market for all kinds of purposes including aviation, maritime, and leisure. To use GPS, a mobile computer user would acquire a receiver, usually in the form of a Compact Flash (CF) card. More recently, receivers are sold as separate devices that can interface with any device that supports the Bluetooth protocol stack. Interestingly, a significant number of mobile phones on the market support Bluetooth, thus offering one scenario for providing location-aware services to mobile phone users.
2. *GLONASS* was developed and deployed by the former USSR in competition to GPS. For a number of years, it was not adequately maintained. However, this situation has changed recently, and GLONASS is currently being overhauled and restored to its former state. There are very few commercial products available that use GLONASS at present.

A third satellite navigation system is scheduled for launch in 2008. *GALILEO* is an initiative by the European Union (EU) that seeks to deliver a similar service to GPS and GLONASS, but with adequate guarantees regarding signal reliability. It is designed for purely civilian and commercial use, and unlike GPS and GLONASS, it is not controlled by defense or military groups. However, the signal broadcast will be compatible with GPS and GLONASS, and it is hoped that receivers that can utilize all three systems will developed.

Cellular Network Techniques

E-911 obligated network operators and, implicitly, telecommunications equipment manufacturers to facilitate the determination of a subscriber's position within an emergency call context. A number of *cellular network techniques* were proposed as a result of ongoing research, and Zhao (2002) provides a useful overview of these. The Third Generation Partnership Project (3GPP) proceeded to standardize on four different techniques for third-generation (3G) UMTS (Universal Mobile Telephone Networks) networks (3GPP, 2005):

1. In *cell-ID*, the geographic coordinates of the base station serving the subscriber are identified. The position of the subscriber must be within the radius of this cell. Though this method is easy to implement, its principle limitation concerns the variability in cell size. Thus the precision with which the subscriber's position is calculated may range from tens to hundreds of meters.
2. *Observed time difference of arrival (OTDOA)* requires the handset to measure the time taken for a signal to arrive from three separate base stations. Hyperbolic curves must be constructed, and their intersection indicates the position of the subscriber. Though computationally expensive, a particular difficulty involves guaranteeing that the subscriber can see three base stations simultaneously. OTDOA is highly susceptible to fading and interference.
3. *Assisted GPS (A-GPS)* involves the handset measuring GPS signals from satellites. Initially, the handset is informed as to where to look for the signals, thus minimizing delay in signal acquisition. The signal measurements are then returned to the appropriate component on the network where the position is calculated. Though increasing power consumption on the device, users can expect position readings comparable with GPS.
4. *Uplink time difference of arrival (UTDOA)* is similar in principle to OTDOA, but in this case, the signals are generated at the handset and measured at a number of base stations. As the geographic positions of the base stations are known, the position of the subscriber can be calculated using hyperbolic trilateration.

With the exception of A-GPS, the accuracy of a position obtained using these techniques is variable and unpredictable. In the case of the cell-ID method, urban areas will have a concentration of base stations so the method may work well. In contrast, the diameter of cells in rural areas may be several kilometers, thus rendering the method ineffective. In the case of OTDOA and UTDOA, accurately measuring the time it takes the signal to travel between the subscriber's handset and surrounding base stations, and vice versa, is essential. Yet the signal may be subject to interference and fading, depending on the vagrancies of the immediate physical environment.

Hybrid Techniques

A scenario can be envisaged where a number of techniques may be combined, with each remedying their respective deficiencies in certain situations. For example, in an urban area, base stations are relatively plentiful, and in certain cases, a number may be deployed in individual streets. Thus techniques like cell-ID, OTDOA, and UTDOA will function reasonably well. In contrast, GPS—and implicitly, A-GPS—may not perform satisfactory, as the high nature of the surrounding buildings, so-called urban canyons, can result in satellites being obscured. In rural areas, the sparsity of base stations may render techniques based on the topology of the cellular network redundant. However, a clear view of the sky is likely, thus GPS and A-GPS should both function satisfactorily.

It should be noted that A-GPS itself could be arguably considered a hybrid technology. However, its close association with and standardization in the telecommunications world result in it being generally considered as a cellular network technique.

The Indoor Scenario

Determining the position of people in an indoor scenario raises particular issues and difficulties. Traditionally, satellite technologies have not operated indoors, as the signal is weak and is subject to additional reflection and fading problems when tracked indoors. A new generation of receivers promises to address this deficiency, with each succeeding generation being incrementally more sensitive. However, the key issues of accuracy and precision remain. This continues to be the case when cellular network techniques are considered, thus making the provision of guarantees concerning the quality of the calculated position exceedingly difficult.

If it is necessary to track a person in an indoor environment with confidence; it is almost essential to consider deploying a dedicated infrastructure, expensive and time-consuming as this may be. However, the required accuracy is a significant determinant. For example, it may be only necessary to track a person to room level. Alternatively, in a museum or art gallery setting, it may be necessary to determine the visitor's position to within one meter so as to determine which artifact is nearest to him or her.

Hightower and Borriello (2001) and Pahlavan, Xinrong, and Makela (2002) provide useful overviews of the issues involved in indoor tracking and positioning. A common approach is to tag the person and place a network of sensors throughout a building. This approach was adopted by Want, Hopper, Falco, and Gibbons (1992) in the pioneering active badge project, and the feasibility of the approach was verified. Systems that use a similar approach today include Cricket (Priyantha, Chakraborty, & Padmanabhan, 2000) and Ubisense (Cadman, 2003). Indeed, given the increased interest in Radio Frequency IDentification (RFID), one can easily envisage a solution involving a fixed network of RFID readers and RFID-tagged personnel.

FUTURE TRENDS

One of the key developments currently taking place concerns the deployment of *satellite-based augmentation systems (SBAS)*. Such systems are a satellite-based implementation of the well-known differential GPS (DGPS) method of improving GPS positions to within a few meters. A number of SBAS systems are being deployed, including the European Ground Navigation Overlay Service (EGNOS) and the Wide Area Augmentation System (WAAS) in the United States. More SBAS satellites are expected to be launched for other areas of the world in the coming years. Two methods for accessing SBAS are of interest. The easiest way is to incorporate an appropriate chip in a GPS receiver. In this way, the position is augmented seamlessly and transparently to the user. A second method involves the Internet, via which SBAS signals can also be broadcast. SISNet (Chen, Toran-Marti, & Ventura-Traveset, 2003) is one example of such a system. Indeed, when A-GPS is reconsidered, it can be seen that integrating this approach with a system such as SISNet is relatively straightforward.

Indoors, the situation is more complex. One approach receiving increasing attention by the research community concerns pseudolites (Wang, 2002). Pseudolites (pseudo-satellites) are placed throughout a building and mimic the GPS signal. Naturally, the pseudolite network should be calibrated for the building in question. However, the important issues of interoperability and standardization—issues that have so far been neglected—must also be addressed.

CONCLUSION

A significant choice of technologies is available for aspiring providers of location-aware services. The required accuracy and precision of the resultant subscriber position is a key determinant of the choice of technology. Attitudes of network operators toward independent small businesses seeking to deploy new services are also of critical importance. It is essential that such operators provide an open and transparent mechanism for accessing subscriber position information. Should the operator adopt an attitude of restricting access or charging excess fees for such information, the potential of location-aware services will be compromised. Overtime, it can be anticipated that a number of mobile phones with integrated GPS and SBAS technologies will be launched on the market. Ultimately, however, it beholds those people designing for mobile users to judiciously consider the merits of the respective positioning technologies in the context of both the application domain and target audience. Only in this way can they be reassured that the needs and expectations of their customers will be addressed.

REFERENCES

Cadman, J. (2003). Deploying commercial location-aware systems. *Proceedings of the Workshop on Location-Aware Computing (held as part of UbiComp)* (pp. 4-6).

Chen, R., Toran-Marti, F., & Ventura-Traveset, J. (2003). Access to the EGNOS signal in space over mobile-IP. *GPS Solutions, 7*(1), 16-22.

Dourish, P. (2004). What we talk about when we talk about context. *Personal & Ubiquitous Computing, 8,* 19-30.

Greenberg, S. (2001). Context as a dynamic construct. *Human-Computer Interaction, 16,* 257-268.

Hightower, J., & Borriello, G. (2001). Location systems for ubiquitous computing. *IEEE Computer, 34*(8), 57-66.

Pahlavan, K., Xinrong, L., & Makela, J.P. (2002). Indoor geolocation science and technology. *IEEE Communications Magazine, 40*(2), 112-118.

Patterson, C. A., Muntz, R. R., & Pancake, C. M. (2003). Challenges in location-aware computing. *IEEE Pervasive Computing, 2*(2), 80-89.

Priyantha, N. B., Chakraborty, A., & Padmanabhan, H. (2000). The cricket location support system. *Proceedings of the 6th ACM International Conference on Mobile Computing and Networking (MOBICOM)* (pp. 32-43).

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *Proceedings of the Workshop on Mobile Computing Systems and Applications* (pp. 85-90). Santa Cruz, CA.

Schmidt, A., Beigl, M., & Gellersen, H.-W. (1999). There is more to context than location. *Computers and Graphics, 23*(6), 893-901.

3GPP. (2005). *3GPP TS 25.305, Technical Specification Group Radio Access Network; Stage 2 Functional Specification of User equipment (UE) positioning in UTRAN (Release 7).*

Vasilakos, A., & Pedrycz, W. (2006). *Ambient intelligence, wireless networking, ubiquitous computing.* Norwood, MA: Artec House, Inc.

Wang, J. (2002). Pseudolite applications in positioning and navigation: Progress and problems. *Journal of Global Positioning Systems, 1*(1), 48-56.

Want, R., Hopper, A., Falco, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

Zhao, Y. (2002). Standardization of mobile phone positioning for 3G systems. *IEEE Communications Magazine, 40*(7), 108-116.

KEY TERMS

GPS: Global positioning system.

OTDOA: Observed time difference of arrival.

Pseudolite: Pseudo satellite.

SBAS: Satellite-based augmentation system.

SISNet (Signal In Space through the Internet): An initiative by the European Space Agency (ESA) to broadcast corrections to the standard GPS signal through the Internet and in real time.

3GPP: Third Generation Partnership Project.

Trilateration: A method of determining the position of an object using the known position of at least three reference points.

UMTS: Universal mobile telephone system.

UTDOA: Uplink time difference of arrival.

Privacy Concerns for Indoor Location-Based Services

Leonardo Galicia Jiménez
CICESE Research Center, Mexico

J. Antonio García-Macías
CICESE Research Center, Mexico

INTRODUCTION

Systems that provide location-based services (LBSs) register all available services in a central entity, so users may subscribe to this entity and thus be able to obtain certain services according to their geographic location. Geographic location information—known as geolocation information—may be known through different mechanisms in manual or automatic ways. Most of these mechanisms imply the use of radiofrequency technologies, through the use of devices such as GPSs, mobile telephones, and others, typically using triangulation techniques to determine the position of the device (and the user carrying it). Most of these systems are implemented for users moving outdoors. Also, geolocation information is under the control of the entity that manages the user's subscription, which motivates concerns regarding how this information is used.

LBS systems require servers where geolocation information is stored, and where objects, attributes, and relations are described in different layers or abstraction levels. This geolocation information enables the representation and visualization of maps, political divisions, roads, electrical networks, buildings, lakes, and so forth. This information is thus stored statically and superposed on a geographical zone, forming a complex map with attributes that can be used to perform queries. These systems commonly use a pull-based paradigm, where users make explicit requests to the server, which responds with geolocation information and related services. For instance, a car driver could request information regarding the closest restaurants, shopping centers, and so on. This example also shows another general characteristic of LBS systems: most of the services are static in nature, meaning that usually the services are in the same place and it is the user who actually moves.

Advances in LBS systems and mobile computing technologies have caught the attention of telephone companies, which seek to provide competitive services to their users and differentiate themselves from their competitors.

Being able to locate artifacts and persons can raise privacy concerns. In fact, these privacy concerns represent one of the most important barriers for the adoption of systems providing location-based services. Thus, effective mechanisms for addressing these privacy concerns should be implemented.

LOCAL MOBILITY

Some of the so-called "knowledge workers" present a high degree of mobility in their daily activities (Belloti & Bly, 1996). Local mobility refers to dynamic patterns of mobility that take place close to the worker's office, or even within a building, when a worker is carrying out her duties, collaborating with colleagues, and so forth. A clear example of this is the kind of work performed in a hospital, as it involves a high degree of mobility of patients, equipment, resources, and personnel within the hospital facilities. Doctors and nurses frequently move in order to carry out their activities; likewise, hospital personnel transport information and equipment through different areas.

Recent studies (Rodríguez & Favela, 2003; Muñoz, Rodriguez, Favela, Gonzalez, & Martinez-Garcia, 2003; Santana et al., 2005) have aided to characterize hospital work in order to design and evaluate different technologies to support processes in that type of environment. As a result, several requirements and needs have been identified for the development of systems that take into account contextual parameters, such as location. Therefore, these needs and requirements also apply for the design and development of LBS systems in hospitals. Among the requirements for information and services identified, related to the location parameter, are:

- **Location of Artifacts:** In many cases, when doctors or nurses finish using medical artifacts, they do not pay attention to returning them to their original place. As a result, when other persons (or even themselves) need these artifacts, they have to invest some time trying to locate them. Checking the availability of some equipment is also a concern.
- **Location of Persons:** During work shifts, medical personnel require locating a specialist to consult on some particular case, or even to get aid in an emergency.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

In this context, the concept of artifacts refers to medical equipment, document, furniture, electronic devices, and any other physical object (static or mobile) that hospital personnel use to carry out or support an activity.

LOCATION OF ARTIFACTS AND PERSONS

In order to provide location-based service discovery, it is first necessary to have some mechanisms to estimate the location of artifacts and persons through some device. Usually, location of devices is basically categorized as device-centered and network-centered. The first one allows the device, through some mechanism, to estimate its own position; that is, it is only the device and no other that can estimate and know its physical location. Meanwhile, network-based mechanisms require that a different entity within the network perform the estimation of a device's physical location; this way, when a device wants to know its location, it has to consult the entity in the network that is in charge of determining it. Some good examples of device-centered systems are RADAR (Bahl & Padmanabhan, 2000), Cricket (Priyantha, Chakraborty, & Balakrishnan, 2000), and AeroScout. In this type of system, a PDA, mobile phone, or some other type of mobile device usually performs the estimation. So, in some way the estimation is person centered, as persons normally carry the devices, but only if the device is fully in control of the calculations for determining the current position; and if the device is turned off for some reason, the location cannot be determined. Some examples of network-centered systems are Active Badge (Want, Hopper, Falcao, & Gibbons, 1992), Ubisense, and Exavera. In these types of systems, objects and persons can carry small devices to aid the network in determining their location.

We think that the network-centered model is more appropriate for the type of scenarios that take place in hospitals, which are our focus for technological development. In a hospital environment some artifacts, such as wheelchairs, stretchers, portable EKG equipment, and others, are good candidates to be located. Moreover, the network-centered model allows the possibility of continuous tracking.

REPRESENTATION OF PHYSICAL SPACES

It is necessary to have a computational model to represent all those artifacts and persons that are moving within a physical space. This model should represent, at least, the physical space, the entities that move within it, as well as those that are static. There are currently different models that allow the representation of physical spaces (Rui, Moreira Rodrigues, & Davies, 2003), including the geometric model, set theory, graphs, and the semantic model. The geometric model includes definitions based on Euclidian geometry, through coordinates in a Cartesian plane; this plane is a direct consequence of cartographic representation, where information and participating entities are superposed on maps, planes, or images. The semantic model offers descriptive information about the geometric areas that represent physical spaces. Under these considerations, both the geometric and semantic models are appropriate for the requirements and technological needs identified in the type of environments that we are interested in (i.e., hospitals). These models have been previously used in the projects and commercial systems mentioned above, namely Cricket, RADAR, Exavera, Ubisense, and Radianse.

PROXIMITY MODEL

A fundamental concept for proximity-based service discovery is, not surprisingly, proximity. The key for a correct association between services and physical spaces is the geographical criterion to be used when services are searched based on their proximity. Two models are widely used for the selection of services: the distance-based model and the scope-based model. In the distance-based model, clients select the services that are within a certain distance from the current position. Given that proximity is a relative value and what is perceived as proximal can vary drastically according to the activities being performed by the client, some mechanism should be present to dynamically change the proximity range.

In a scope-based model, each service is associated with a scope that explicitly represents the context of use of the service within a physical space. The client selects those services whose scopes include the location of the requesting client; that is, a client can discover services if it is inside a certain scope, as well as the services. The main characteristic of this model is that the correlation between context and proximity is assured. When services are discovered, no matter what their distance, they have a high probability of being relevant for the requesting client.

We consider that the scope-based model better suits our needs, mainly because it allows the representation of physical sub-spaces as geometric shapes; this is very adequate for indoor environments such as hospitals where the definition of rooms, working areas, and so on is very useful.

DEFINITION OF SERVICES

A service is an entity that can be used by a person, a computer program, or any other entity (Johansen, 1999); examples of services are files, a storage device, a printer, a server. When service discovery is performed within a physical space or scope, it is not enough to know the available services, but also the persons within it, in order to be able to interact with them. Thus, it is convenient to consider artifacts and persons as services associated to a given scope; it should be noted that some of these entities will not strictly provide a service, but it is necessary to know their attributes (type, role, etc.) in order to know what kind of entities there are and where they are in a certain space.

USER IDENTIFICATION

In a ubiquitous computing environment, the search for services responds to contextual variables such as the location of the user and its identification. While the location allows obtaining those services near the user, the identification allows the user to access them. This way, a user could make an anonymous request for available services within a scope, and she may obtain a list of them. However, it could be possible that more services would exist; this is because some of those services may require additional information about the user (such as her role, an ID, etc.) in order to be possible to discover them. This way, services are discovered not only based on their location or proximity, but also based on who is requesting them.

CONTROL OF ACCESS TO LOCATION INFORMATION

In spite of the benefits that ubiquitous computing environments offer, one of the main barriers for the adoption of related technologies is the privacy concerns of users. A clear example is shown in the results obtained by Intel Research Berkeley through a series of interviews and scenarios (Barkhuus & Dey, 2003) where users manifested concerns regarding being located. Recently the IETF, through the Geopriv working group, has been studying privacy issues that arise when the geographical information of people and resources is used. The focus of the group, as stated in their charter, is "to assess the authorization, integrity and privacy requirements that must be met in order to transfer such information, or authorize the release or representation of such information through an agent." Then, attention is paid to presence and geospatial information, commonly used in instant messaging systems, location-based services, and others. Until now, the Geopriv working group has generated a couple of RFCs (requests for comments) and recommendations aimed at proposing a standard that guarantees the privacy of users. Much of this work is currently under review, but rules and mechanisms have already been defined (represented in XML formats); these control when, to whom, in what place, and under what circumstances geolocation information can be released. The Geopriv group uses formats and architectures previously approved by the IETF, for instance XMPP (Extensible Messaging and Presence Protocol) or SIMPLE (SIP for Instant Messaging and Presence Leveraging Extensions).

The user should be able to define when (day and time), to whom, where (place), and under what circumstances (status of the user) location information can be released. The mechanisms for this are defined by the Geopriv group and can be easily applied to artifacts, not for protecting their privacy, as this would not make much sense, but for extending service discovery and indicating who and under what circumstances they can be discovered. This way, artifacts and persons (viewed more generally as services) can be discovered based not only on their location, but also taking into account who is requesting them using control policies for service discovery.

These policies are inspired mainly by RFC 3669 of the Geopriv working group, as well as by the policy and common policy working drafts of that same group. RFC 3669 deals with the authorization, integrity, and privacy involved in releasing information about the location of users. The Geopriv group is currently exploring how to represent information about location and presence, as well as how to protect this information.

In its Common Policy draft, the Geopriv group defines the base mechanisms for delivering location information in presence messages; these mechanisms, which allow access control form information regarding presence and location of users, can be extended and translated easily to other application domains. These mechanisms define an XML document that represents policies associated to an entity. Requests from entities contain policy rules, and these are checked against the policies defined in the entity that receives the request; if one or more of the rules match, then the location and presence information are released, else it is denied.

Three sections compose each rule: conditions, actions, and transformations. The conditions section defines all the restrictions that should be satisfied by an entity in order for it to obtain the requested information. The actions section is a set of processes that the user requests an entity to perform; these actions have not been defined by Geopriv and are meant to be defined at the application (and not at the user) level. The transformation section indicates those modifications that should be made to the location information before being released. For instance, even if a user complies with all the rules imposed by an entity, a transformation could be imposed to reduce the precision of the geographic location, or to indicate the floor where the entity is located but without saying in what room it is.

CONCLUSION

We have presented the design considerations that should be taken into account for the design of ubiquitous distributed applications that allow the discovery of service based on their physical proximity and considering important privacy concerns. The use of privacy policies, as proposed by the IETF Geopriv working group, can be applied in order to restrict the way in which services are discovered and the potentially sensitive information is disclosed.

REFERENCES

Bahl, P., & Padmanabhan, V.N. (2000). RADAR: An in-building RF-based user location and tracking system. *IEEE INFOCOM, 2,* 775-784.

Barkhuus, L., & Dey, A. (2003). Location-based services for mobile telephony: A study of users' privacy concerns. *Proceedings of INTERACT 2003, the 9th IFIP TC13 International Conference on Human-Computer Interaction.*

Belloti, V., & Bly, S. (1996). Walking away from the desktop computer: Distributed collaboration and mobility in a product design team. *Proceedings of CSCW* (pp. 209-218). ACM Press.

Bettstetter, C., & Renner, C. (2000). A comparison of service discovery protocols and implementation of the service location protocol. *Proceedings of the 6th EUNICE Open European Summer School.*

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research.* Dartmouth Computer Science Technical Report TR2000-381, USA.

Dahlbom, B., & Ljungberg, F. (1998). Mobile informatics. *Scandinavian Journal of Information Systems, 10*(1-2), 227-234.

Dey, A., & Abowd, G. (2000). Towards a better understanding of context and context-awareness. *Proceedings of the Workshop on the What, Who, Where, When and How of Context-Awareness at CHI.*

Dix, A. (2000). Exploiting space and location as a design framework for interactive mobile systems. *ACM Transactions on Computer-Human Interactions,* 285-321.

Droms, R. (1997). *Dynamic host configuration protocol.* RFC 1541, Internet Engineering Task Force (IETF).

Guttman, E., Perkins, C., Veizades, J., & Day, M. (1998). *Service location protocol, version 2.* RFC 2608, Internet Engineering Task Force (IETF).

Harter, A., Hopper, A., Steggles, P., Ward, A., & Webster, P. (2002). The anatomy of a context-aware application. *Wireless Networks, 8*(2-3), 187-197.

Helal, S. (2002). Standards for service discovery and delivery. *IEEE Pervasive Computing, 1*(3), 95-100.

Hightower, J., Boriello, G., & Want, R. (2002). *SpotON: An indoor 3D location sensing technology based on RF signal strength.* Technical Report 2000-02-02, University of Washington, USA.

Hodes, T., Katz, R., Servan-Schreiber, E., & Rowe, L. (1997). Composable ad-hoc mobile services for universal interaction. *Proceedings of the 3rd ACM/IEEE International Conference on Mobile Computing* (pp. 1-12).

Johansen, T. (1999). *Jini architectural overview.* White Paper, Sun Microsystems, USA.

Microsoft. (n.d.). *Universal plug and play device architecture reference specification, version 1.0.* Technical Report, Microsoft Corporation, USA.

Muñoz, M., Rodriguez, M., Favela, J., Gonzalez, V.M., & Martinez-Garcia, A.I. (2003). Context-aware mobile communication in hospitals. *IEEE Computer, 36*(8), 60-67.

Priyantha, N.B., Chakraborty, A., & Balakrishnan, H. (2000). The Cricket location-support system. *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking* (pp. 32-43). ACM Press.

Rodríguez, M., & Favela, J. (2003). Autonomous agents to support interoperability and physical integration in pervasive environments. *Proceedings of the Atlantic Web Intelligence Conference* (AWIC 2003) (pp. 278-287). Berlin: Springer-Verlag.

Rui, J., Moreira Rodrigues, A., & Davies, N. (2003). The AROUND architecture for dynamic location-based services. *Mobile Networks and Applications, 8*(4), 377-387.

Santana, P., Castro, L.A., Preciado, A., Gonzalez, V.M., Rodríguez, M.D., & Favela, J. (2005). Preliminary evaluation of Ubicomp in real working scenarios. *Proceedings of the 2nd Workshop on Multi-User and Ubiquitous User Interfaces* (MU3I).

Schilit, B., Adams, N., Want, R. (1994). Context-aware computing applications. *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications* (pp. 85-90).

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The Active Badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

Weiser, M. (1991). The computer for the twenty-first century. *Scientific American, 265*(3), 94-104.

KEY TERMS

Context: Any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between the user and an application, including the user and the application themselves.

Context-Aware Computing: Augments computers with sensors and actuators in order to achieve a better understanding of, and interaction with, the physical environment. Such systems collect sensor data, build a model of the environment or world model, and use the model to provide more useful and intuitive services to users by triggering actuators or automating tasks (Schilit et al., 1994).

Local Mobility: Describes the need of some workers to displace and move around their offices or premises to conduct their work (Belloti & Bly, 1996).

Privacy: The ability of an individual or group to stop information about themselves from becoming known to people other than those they choose to give the information to. Privacy is sometimes related to anonymity, although it is often most highly valued by people who are publicly known. Privacy can be seen as an aspect of security—one in which trade-offs between the interests of one group and another can become particularly clear.

Service Discovery Protocol: Computational mechanisms for the provision of adequate support to dynamic changes in the network regarding the services it offers. These mechanisms allow clients to connect to the network and be able to search for a particular service, and also allow entities providing services to announce their capabilities; all these without previous configuration.

Service or Resource: An entity that can be used by a person, by a computer program, or by another service. For instance: a file, a storage device, a computer (Johansen, 1999).

Ubiquitous Computing: A concept proposed by Mark Weiser (1991). Refers to the use of computers embedded in the physical space of the user. Through these computer-saturated spaces, the user has some of them available to support his or her activities (Schilit, Adams, & Want, 1994). These computers are small devices with ample processing power, and given their small size they are used without the user being aware of them.

Protocol Analysis for the 3G IP Multimedia Subsystem

Muhammad Tanvir Alam
Bond University, Australia

INTRODUCTION

In the past few years, the evolution of cellular networks has reflected the success and growth the Internet has experienced in the last decade. This leads to networks where IP connectivity is provided to mobile nodes. The result is third-generation (3G) networks where IP services such as voiceover IP (VoIP) and instant messaging (IM) are provided to mobile nodes (MNs) in addition to connectivity. IP multimedia subsystem (IMS) is a new framework, basically specified for mobile networks, for providing Internet protocol (IP) telecommunication services. It has been introduced by the Third-Generation Partnership Project (3GPP) in two phases (release 5 and release 6) for Universal Mobile Telecommunications System (UMTS) networks. 3GPP was born in 1998 as a collaboration agreement between a number of regional telecommunication standards bodies, known as *organizational partners.* The current 3GPP organizational partners are:

- ARIB (Association of Radio Industries and Business) in Japan,
- CCSA (China Communications Standards Associations) in China,
- ETSI (European Telecommunications Standards Institute) in Europe,
- Committee T1 in the United States of America
- TTA (Telecommunications Technology Association) of Korea, and
- TTC (Telecommunication Technology Committee) in Japan.

Besides the organizational partners, *market representation partners* (the UMTS Forum, 3G Americas, the IPv6 Forum, the Global Mobile Suppliers Association, etc.) provide the partnership with market requirements. 3GPP maintains an up-to-date Web site at http://www.3gpp.org.

3GPP working groups do not provide standards. Instead, they produce technical specifications (TSs) and technical reports (TRs). 3G PP Release 5 contains the first version of the IMS. 3GPP release 6 contains enhancements to the IMS. The Third Generation Partnership Project 2 (3GPP2) was born to evolve North American and Asian cellular networks based on ANSI/TIA/EIA-41 standards and CDMA2000 radio access into a third-generation system. An IP multimedia framework was later introduced by 3GPP2 as the Multimedia Domain (MMD) for third-generation Code Division Multiple Access 2000 (CDMA2000) networks, and finally harmonized with IMS. IMS aims to use Internet protocols. The 3GPP and 3GPP2 established collaboration with the IETF to make sure that the protocols developed there meet their requirements. In this article, we aim to discuss a few potential areas of the IMS that could be improved to provide better quality of service (QoS).

BACKGROUND

The IMS is the technology that will merge the Internet (packet switching) with the cellular world (circuit switching). It will make Internet technologies, such as the Web, e-mail, instant messaging, presence, and videoconferencing- available nearly everywhere.

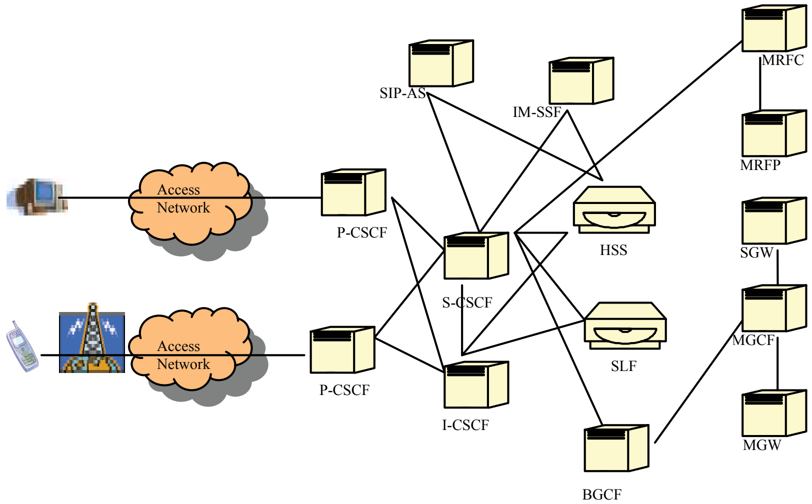
In brief the IMS concept was introduced to address the following network and user requirements:

- Deliver person-to-person, real-time, IP-based multimedia communications (e.g., voice or video-telephony) as well as person-to-machine communications (e.g., gaming service).
- Fully integrate real-time with non-real-time multimedia communications (e.g., live streaming and chat).
- Enable different services and applications to interact (e.g., combined use of presence and instant messaging).
- Easy user setup of multiple services in a single session or multiple simultaneous synchronized sessions.

Figure 1 depicts an overview of the IMS architecture. The definitions and functions of the common nodes included in the IMS are furnished in the Key Terms section. There are plenty of ways to improve the existing infrastructure and protocols in the IMS. Previously, the performance of the critical protocol SIP (Session Initiation Protocol, Rosenberg et al., 2002) in an IMS environment had never been evaluated. The signaling overhead reaches its peak when a massive number of IMS terminals joins the network at the same time. The minimal discovery time of different CSCFs are crucial for system performance.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. 3GPP IMS architecture overview



Session Establishment Scenario for a Mobile Terminal

Every mobile node must register with the visited network in IMS. Re-registration takes place once the timeout occurs. The SIP INVITE request is sent from the UE (user equipment) to S-CSCF#1 (serving call session control function) by the procedures of the originating flow to initiate a session between two nodes. This message may contain the initial media description in the SDP (session description protocol). S-CSCF#1 performs an analysis and passes the request to I-CSCF#1 (interrogating CSCF) and so on. Thus the intermediate nodes analyze and forward the request to the next node until it reaches the destination node. The detail of IMS SIP session set up procedures with MIPv6 can be found in Technical Specification 23.228 of IP Multimedia Subsystems.

If a mobile terminal moves away from its current visited network, it needs to send a binding update (BU) message to the corresponding node. It may move away during the session set up. The issue of sending the BU to achieve better mobility management needs to be addressed thoroughly.

In the existing scenario, the mobile node sends the BU to the corresponding node after the session is set up. This implies that traffic will be routed through the HA (home agent) before being routed directly to the MN (mobile node), even if for a limited amount of time. This can have implications on quality of service, since quality of service (QoS) is initially established only for the route from the MN to the HA and to the CN (correspondent node), whereas QoS for the optimized route is not established.

Presence Service in the IMS

Presence is one of the basic services that is likely to become omnipresent in IMS. It is the service that allows a user to be informed about the reachability, availability, and willingness of communication of another user.

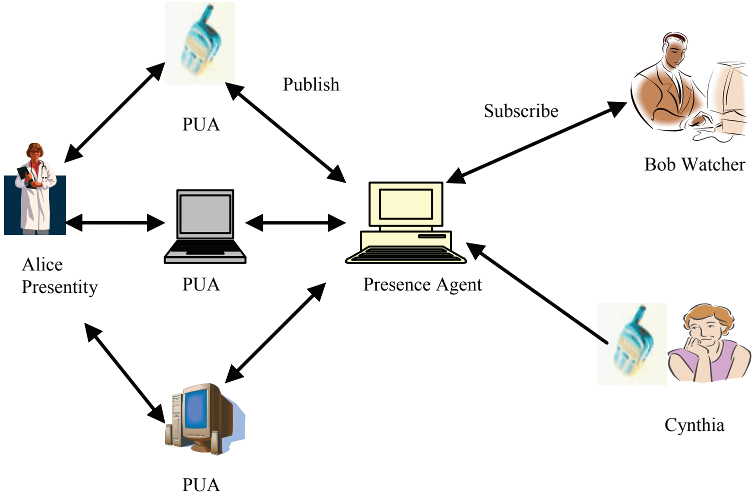
The presence framework defines various roles as shown in Figure 2. The person who is providing presence information to the presence service is called a presence entity, or for short a presentity. In the figure, Alice plays the role of a presentity. The presentity is supplying presence information such as status, capabilities, communication address, and so forth. A given presentity has several devices known as presence user agents (PUAs) which provide information about her presence. All PUAs send their pieces of information to a presence agent (PA). A presence agent can be an integral part of a presence server (PS). A PS is a functional entity that acts as either a PA or as a proxy server for SUBSCRIBE requests. Figure 3 also shows two watchers: Bob and Cynthia. A watcher is an entity that requests (from the PA) presence information about a presentity or watcher information about his/her watchers. A subscribed watcher asks to be notified about future changes in the presentity's presence information, so that the subscribed watcher has an updated view of the presentity's presence information.

3GPP defined in 3GPP TS 23.141 provides the architecture to support the presence service in the IMS. The architecture indicates that the flow of messages will be massive for a large amount of publishers and watchers joining an IMS system.

P-CSCF Discovery

P-CSCF discovery is the procedure by which an IMS terminal obtains the IP address of a P-CSCF. This is the P-CSCF that acts as an outbound/inbound SIP proxy server toward the IMS terminal (i.e., all the SIP signaling sent by or destined for the IMS terminal traverses the P-CSCF). P-CSCF discovery may take place in two different ways:

Figure 2. SIP presence architecture



1. integrated into the procedure that gives access to the IP-CAN (IP connectivity access network), or
2. as a stand-alone procedure.

The integrated version of P-CSCF discovery depends on the type of IP connectivity access network. If IP-CAN is a GPRS (general packet radio service) network, once the GPRS attach procedures are completed, the terminal is authorized to use the GPRS network. Then the IMS terminal does a so-called activate PDP context procedure. The main goal of the procedure is to configure the IMS terminal with an IPv6 address, but in this case the IMS terminal also discovers the IPv6 address of the P-CSCF to which to send SIP requests.

The stand-alone version of the P-CSCF discovery is based on the use of DHCPv6 (Dynamic Host Configuration Protocol for IPv6) specified in RFC 3315 by Droms et al. (2003) and DNS (Domain Name System, specified in RFC 1034 by Mockapetris, 1987). A suitable procedure is required in order to identify the faster mechanism to discover P-CSCF in IMS.

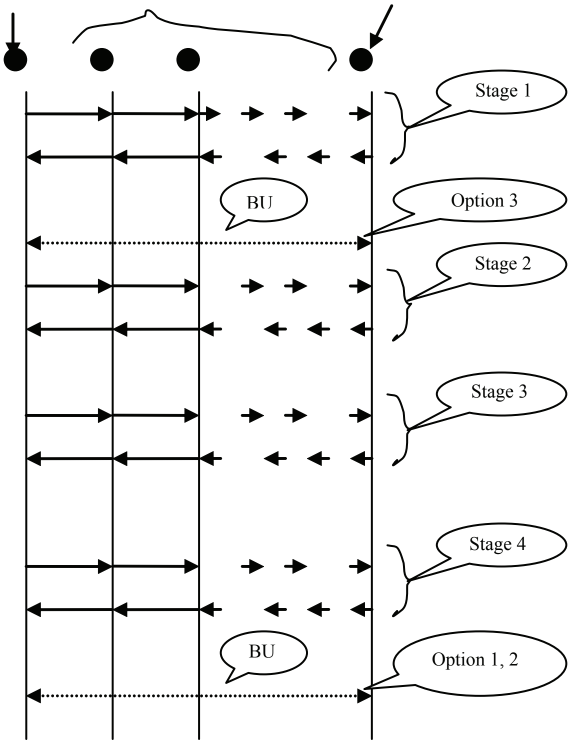
CHALLENGES OF THE IMS

The abovementioned aspects could be potentially reviewed. Following are some proposed approaches to improve some of the IMS services.

Session Establishment Schemes in Mobile Environment

An architecture of mobility and QoS support is provided by Shou-Chih Lo, Lee, Wen-Tsuen, and Jen-Chi (2004). Nakajima, Dutta, Das, and Schulzrinne (2003) depicted a handoff delay analysis for SIP-based mobility in IPv6. The whole procedure of session set up in IMS may be divided into four stages as shown in Figure 3. Stage one includes sending of an INVITE message from source to destination and getting a response from the destination back to the source. Stages two, three, and four can be described in the same manner

Figure 3. Three options for IMS session set up



as Response, Reservation, and ACK messages respectively. Note that a session set up may fail anytime due to the different processing complexity. The destination node may send a BUSY message in response to the INVITE request. The message may also be corrupted or lost in the intermediate nodes which raises the possibility for a session to fail at any stage during the period of establishment.

The basic scenario (option 1) (Johnson, Perkins, & Arkko, 2004) is that the MN receives packets from the CN tunneled though the HA, and initiates the route optimization procedure. This implies that traffic will be routed through the HA before being routed directly to the MN, even if for a limited amount of time. This can have implications on quality of service, since QoS is initially established only for the route from the MN to the HA and to the CN, whereas QoS for the optimized route is not established.

A second scenario (option 2) introduces an optimization where the MN sends a BU to the CN immediately after setting up the SIP call, before any traffic is received from the CN. This option was mentioned by Faccin, Lalwaney, and Patil (2004). This requires slight modifications to the implementation of the MN, but benefits from route optimization from the beginning of the communication.

The main difference between options 1 and 2 lies in the data transfer. MN may start sending the data via HA immediately after the session is set up, before the BU has been sent in Option 1. MN waits till the BU has been sent before it starts to send any data in Option 2.

We propose an additional optimization (option 3) by sending the BU message in parallel while the SIP session is still being set up. For example, after the first round trip (which includes the INVITE message reaching the destination and coming back to the source) of messages of a session set up, the source may initiate (by forking) a BU for the destination and try to ensure QoS. Alternatively, the BU message could be initiated after the second roundtrip of a session set up progress. This way, the MN can immediately start to send data once the session has been set up. However, the overhead would be high if a session fails to set up for variable reasons.

Note that options 1 and 2 are applicable only for a successful session set up, while option 3 is applicable to both successful and unsuccessful session set ups. In options 1 and 2, the BU is sent only after the session is set up, while the BU is sent in parallel in option 3. The idea is to compare and contrast the different scenarios with an appropriate algorithm. The overhead and the delay of the schemes need to be identified.

Acquiring an IP for P-CSCF

In GPRS, the IMS terminal first undertakes a set of procedures, globally known as GPRS attach procedures, in order to acquire an IP for P-CSCF. These procedures involve several nodes, ranging from the SGSN to the HLR and the GGSN. Once these procedures are complete, the terminal sends an activate PDP context request message to the SGSN requesting connection to an IPv6 network. The message includes a request for connectivity to a particular APN (access point name) and packet connection type. The APN identifies the network to connect to and the address space where the IP address belongs. In the case of an IMS terminal, the APN indicates a desired connection to the IMS network and the connectivity type indicates IPv6. The SGSN, depending on the APN and the type of network connection, chooses an appropriate GGSN. The GGSN is responsible for allocating IPv6 addresses. In the case of the IMS, the GGSN does not provide the terminal with an IPv6 address belonging to the IMS address space. Instead, the GGSN provides the terminal with a 64-bit IPv6 prefix and includes it in a create PDP context response message. The SGSN transparently forwards this IPv6 prefix in an active PDP context accept. When the procedure is completed, the IMS terminal has got a 64-bit IPv6 prefix. The terminal is able to choose any 64-bit IPv6 suffix. Together they form a 128-bit IPv6 address—that is, the IPv6 address that the terminal will use for its IMS traffic. During this GPRS attach procedure, the terminal also discovers the IPv6 address of the P-CSCF in a similar fashion from GGSN.

The stand-alone version of discovering P-CSCF depends on DCHP and DNS. In DHCP the terminal does not need to know the address of the DHCP server, because it can send its DHCP messages to a reserved multicast address. In some configurations a DHCP relay may be required to relay DHCP messages to an appropriate network, although the presence of the DHCP relay is transparent to the terminal.

Once the terminal has connectivity to the IP-CAN, the IMS terminal sends a DHCPv6 information request where it requests the DHCPv6 Options for SIP servers (specified by Schulzrine & Volz, 2003). In the case of the IMS, the P-CSCF performs the role of an outbound/inbound SIP proxy server, so the DHCP server returns a DHCP reply message that contains one or more domain names and/or IP addresses of one or more P-CSCFs.

At the discretion of the IMS terminal implementation, there are two possible ways in which the IMS terminal can specify the request for the DHCPv6 Option for SIP servers:

1. The IMS terminal requests the SIP server's domain name list option in the DHCPv6 information-request message. The DHCPv6 reply message contains a list of the domain names of potential P-CSCFs. The IMS terminal needs to resolve at least one of these domain names into an IPv6 address. A query response dialog with DNS resolves the P-CSCF domain name, but prior to any DNS interaction, the IMS terminal also needs to get the address of one or more DNS servers

to send its DNS messages. To resolve this problem, the DHCP information-request message not only contains a request for the option for SIP servers, but also includes a request for the DNS recursive name server option. The DHCPv6 reply message contains a list of IPv6 addresses of DNS servers, in addition to the domain name of the P-CSCF. Then, the IMS terminal queries the just learned DNS server in order to resolve the P-CSCF domain name into one or more IPv6 addresses. The procedures to resolve a SIP server into one or more IP addresses are standardized in RFC 3263 (Rosenberg & Schulzrine, 2002).

2. The alternative consists of the IMS terminal requesting the SIP server's IPv6 address list option in the DHCPv6 information-request message. The DHCP server answers in a DHCP reply message that contains a list of IPv6 addresses of the P-CSCF allocated to the IMS terminal. In this case, no interaction with DNS is needed, because the IMS terminal directly gets one or more IPv6 addresses.

Eventually, the IMS terminal discovers the IP address of its P-CSCF and can send SIP signaling to its allocated P-CSCF. However, these procedures do not mention which one is efficient in which environment. For example, if P-CSCF is located in the home network and the CN (IMS terminal) in the visited network, which would be the quicker method for P-CSCF discovery? The performance of network access connectivity in IMS was never analyzed before. An efficient algorithm is needed for an IMS terminal to promptly select the faster method in order to achieve the IP address of a P-CSCF to send SIP messages.

Presence Service during Heavy Traffic

In order to reduce heavy message flows in the IMS presence service, the IETF has created a number of concepts described as follows.

1. Partial notification is one mechanism on which IETF engineers are working to reduce the amount of presence information transmitted to watchers. The mechanism defines a new XML body that is able to transport partial or full state. Thus, the document size is reduced at the cost of information transmitted.
2. An event-throttling mechanism allows a subscriber to an event package to indicate the minimum period of time between two consecutive notifications. So, if the state changes rapidly, the notifier holds those notifications until the throttling timer has expired, at which point the notifier sends a single notification to the subscriber. However, with this mechanism the watcher does not have a real-time view of the subscription state information.
3. Compression of SIP messages is another technique to minimize the amount of data sent on low-bandwidth access. RFC 3486 by Camarillo (2003), RFC 3321 by Hannu et al., 2003a), and RFC 3320 by Hannu et al. (2003b) define signaling compression mechanisms. Usually these algorithms substitute words with letters. The compressor builds a dictionary that maps the long expressions to short pointers and sends this dictionary to the de-compressor. However, the frequency of data transmission is not reduced in such techniques.

Clearly each of the abovementioned works has limitations and tradeoffs. The lifetime of a watcher subscription time has not received any attention so far. An optimal watcher registration time procedure to allow Proxy-CSCF to reassign UE (user equipment) in IMS needs to be considered. Every time, the UE/IMS watcher needs to re-subscribe when its timer (which is kept shorter than the subscription timer in the network) expires. If the UE does not re-register, any of its active sessions are deactivated in IMS. On the other hand, de-registration is accomplished by a registration with an expiration time of zero seconds. A forced de-registration from the network (PA) may occur in case of data inconsistency at node failure. The constant time set may create a bottleneck because of excessive message flow in the network. Especially, if an IMS watcher watches many presentities and if the watcher-subscription-time is not set carefully, it will be notified of any changes made in its presentity list. Both long and short lifetime will introduce overhead in a number of messages and cache respectively. Thus an optimal procedure to set the timer of the watcher subscription lifetime for the IMS node is desirable.

FUTURE TRENDS

The IMS has yet to undergo further refinement to overcome all its shortcomings. New algorithms to reduce latency in session set ups in mobile environments, to balance load efficiently for heavy traffic PS, and to make faster discovery of P-CSCF are required. Besides, there are other potential areas that need to be addressed. IMS provides SIP-PSTN inter-working services. The traditional audio calls in SIP-PSTN inter-working focus on audio-only calls. Although, the video services are included in the inter-working, significant work is required in the encoding and gateway architecture to facilitate video streaming in the IMS. One other problem with the IMS is that SIP does not offer a mechanism for a UA (user agent) to indicate that all proxies in the path must use a transport protocol that implements end-to-end congestion control—that is, SIP derives from the fact that any proxy can change the transport protocol from TCP (transmission control protocol) to UDP (user datagram protocol), SCTP (stream control transmission protocol), or others and vice versa. UDP is notorious for not offering congestion control, which may affect the Instant Messaging services in the IMS. The IETF engineers are working on an extension to SIP that will allow a UA to request proxies to use a transport protocol that supports end-to-end congestion control. Session-based instant messaging may be a solution to this problem that uses MSRP (message session relay protocol, by Campbell, Mahy, & Jennings, 2005). MSRP is a simple text-based protocol whose main characteristic is that it runs over transport protocols that offer congestion control. However, the complete behavior of MSRP relays is still not well defined. The future releases of the IMS should refer to the abovementioned areas at least.

CONCLUSION

This research depicts some of the potential approaches to some of the shortcomings of the IMS. One of the reasons for creating the IMS was to provide the quality of service required for enjoying, rather than suffering, real-time multimedia sessions. Thus its drawbacks are essential to be overcome in the near future. The proposed Option 3 might reduce delay in session establishment while the IMS terminals are mobile. However, the overhead needs to be tested if a session fails to set up in the first trial. One potential solution for reducing load in the IMS presence service may be to optimize the watcher subscription time—that is, a watcher should not be able to watch its presentities' infinite amount of time during heavy traffic to reduce heavy message flow. Again, the scheme needs to be tested thoroughly in a suitable environment.

REFERENCES

Camarillo, G. (2003). *Compressing the Session Initiation Protocol (SIP).* RFC 3486, Internet Engineering Task Force.

Campbell, B., Mahy, R., & Jennings, C. (2005). The Message Session Relay Protocol MSRP. Retrieved from draft-ietf-simple-message-sessions-10.txt

Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., & Carney, M. (2003, July). *Dynamic Host Configuration Protocol for IPv6 (DHCPv6).* RFC 3315, Internet Engineering Task Force.

Faccin, S. M., Lalwaney, P., & Patil, B. (2004). IP multimedia services: Analysis of mobile IP and SIP interactions in 3G networks. *IEEE Communications Magazine, 8*(1), 113-118.

Hannu, H., Christoffersson, J., Forsgren, S., Leung, C., Liu, Z., & Price, R. (2003a, January). *Signaling Compression (SigComp)—Extended operations.* RFC 3321, Internet Engineering Task Force.

Hannu, H., Rosenberg, J., Bormann, C., Christoffersson, J., Liu, Z., & Price, R. (2003b, January). *Signaling Compression (SigComp).* RFC 3320, Internet Engineering Task Force.

Johnson, D. B., Perkins, C., & Arkko, J. (2004). *Mobility support in IPv6.* RFC 3775, IETF Mobile IP Working Group.

Mockapetris, P. V. (1987, November). *Domain names—Concepts and facilities.* RFC 1034, Internet Engineering Task Force.

Nakajima, N., Dutta, A., Das, S., & Schulzrinne, H. (2003). Handoff delay analysis and measurement for SIP based mobility in IPv6. *Proceedings of the International Conference on Communications (ICC '03)* (*V*ol. 2, pp. 11-15).

Rosenberg, J., & Schulzrine, H. (2002). *Session initiation protocol (SIP): Locating SIP servers.* RFC 3263, Internet Engineering Task Force.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., & Schooler, E. (2002). *SIP: Session initiation protocol.* RFC 2543, Internet Engineering Task Force.

Schulzrine, H., & Volz, B. (2003). *Dynamic Host Configuration Protocol (DHCPv6) options for Session Initiation Protocol (SIP) servers.* RFC 3319, Internet Engineering Task Force.

S.-C. Lo, Lee, G., Wen-Tsuen, C., & Jen-Chi, L. (2004). Architecture for mobility and QoS support in all-IP wireless networks. *IEEE Journal on Selected Areas in Communications, 22*(4), 691-705.

3GPP. (2005.). *Presence service; architecture and functional description; stage 2.* TR 23.141, Third Generation Partnership Project.

3GPP. (2002). *Technical specification group services and system aspects, IP Multimedia Subsystem (IMS)—Stage 2 (release 5).* TS 23.228 v5.6.0 (2002-2009).

3GPP TSG SSA. (2004). *IP Multimedia Subsystem (IMS)—Stage 2 (release 6).* TS 23.228 v. 6.6.0 (2004-2006).

KEY TERMS

Breakout Gateway Control Function (BGCF): A session initiation protocol server that includes routing functionality based on telephone numbers.

Call/Session Control Function (CSCF): A session initiation protocol (SIP) server that processes SIP signaling in the IP multimedia subsystem. There are three types of CSCFs depending on the functionality they provide.

Home Subscriber Server (HSS): Contains all the user-related subscription data required to handle multimedia sessions. These data include, among other items, location information, security information (including both authentication and authorization information), user profile information, and the S-CSCF allocated to the user. The SLF (subscription location function) is a simple database that maps users' addresses to HSSs. Both the HSS and the SLF implement the Diameter protocol.

Interrogating-CSCF (I-CSCF): Provides the functionality of a SIP proxy server. It also has an interface to the SLF (subscriber location function) and HSS (home subscriber server). This interface is based on the diameter protocol. The I-CSCF retrieves user location information and routes the SIP request to the appropriate destination, typically an S-CSCF.

IP Multimedia Services Switching Function (IM-SSF): Acts as an application server on one side, and on the other side, it acts as an SCF (service switching function) interfacing the gsmSCF (GSM service control function) with a protocol based on CAP (CAMEL application part).

Media Gateway (MGW): Interfaces the media plane of the PSTN (public-switched telephone network) or CS (circuit-switched) network. On one side the MGW is able to send and receive IMS media over the real-time protocol (RTP). On the other side the MGW uses one or more PCM (pulse code modulation) time slots to connect to the CS network. Additionally, the MGW performs transcoding when the IMS terminal does not support the codec used by the CS side.

Media Gateway Control Function (MGCF): Implements a state machine that does protocol conversion and maps SIP to either ISUP (ISDN user part) over IP or BICC (bearer independent call control) over IP. The protocol used between the MGCF and the MGW is H.248.

Media Resource Function (MRF): Provides a source of media in the home network. It is further divided into a signaling plane node called the MRFC (media resource function controller) and a media plane node called the MRFP (media resource function processor). The MRFC acts as a SIP User Agent and contains a SIP interface towards the S-CSCF. The MRFC controls the resources in the MRFP via an H.248 interface. The MRFP implements all the media-related functions.

Proxy-CSCF (P-CSCF): The first point of contact between the IMS terminal and the IMS network. All the requests initiated by the IMS terminal or destined to the IMS terminal traverse the P-CSCF. This node provides several functions related to security. The P-CSCF also generates charging information toward a charging collection node. An IMS usually includes a number of P-CSCFs for the sake of scalability and redundancy. Each P-CSCF serves a number of IMS terminals, depending on the capacity of the node.

Serving-CSCF (S-CSCF): A SIP server that performs session control. It maintains a binding between the user location and the user's SIP address of record (also known as public user identity). Like the I-CSCF, the S-CSCF also implements a diameter interface to the HSS.

Signaling Gateway (SGW): Performs lower-layer protocol conversion.

SIP Application Server (SIP AS): The AS is a SIP entity that hosts and executes IP multimedia services based on SIP.

Protocol Replacement Proxy for 2.5 and 3G Mobile Internet

Victor Khashchanskiy
First Hop Ltd., Finland

Andrei Kustov
First Hop Ltd., Finland

Jia Lang
Nice Business Solutions Finland, Finland

INTRODUCTION

Providing mobile Internet access in GPRS and UMTS networks is not an easy task. The main problem is in rather challenging network conditions (Inamura, Montenegro, Ludwig, Gurtov, & Khafizov, 2003). Latency in these networks could be an order of magnitude higher than in wired networks, with round-trip time (RTT) reaching up to one second. Moreover, there occur delay spikes in the network, when latency can exceed average RTT several times (Gurtov, 2004). Furthermore, in wireless networks, the risk of experiencing packet losses is considerably higher in comparison to that in wired networks. This is because packets can easily be lost due to corruption, either during deep fading leading to burst losses, or cell re-selections, resulting in a link black-out condition. Such characteristics of wireless cellular networks significantly affect performance of the principal Internet protocol—TCP—as it was designed to work in conditions of low-latency reliable networks.

TCP assumes that all segment losses indicate congestion as they are traditionally (i.e., in wired links) caused by buffer overflows in routers. If losses are detected, in addition to retransmitting the lost packet, TCP adjusts the values of sending window size and retransmission timeout (RTO) in order to slow down transmission. Packet losses or even long enough delays can lead to TCP timeouts. However, in both cases the timeouts are not caused by congestion, hence the basic working assumption for TCP is incorrect, and consequently the countermeasures are also not optimal. TCP flow control is achieved by complex mechanisms trying to probe for a data rate as high as possible but backing off as soon as congestion occurs. A TCP sender adapts its use of bandwidth based on feedback from the receiver. The high latency characteristic of cellular networks implies that TCP adaptation is correspondingly slower than in wired networks with shorter delays. Similarly, delayed acknowledgements exacerbate the perceived latency on the link.

The central performance issues of TCP in wireless cellular networks lie in the inability to correctly detect the nature of the error, and so it is incapable of responding in an appropriate manner (Tsaoussidis & Matta, 2001). In addition, the protocol lacks efficient monitoring of the network conditions, rapid window size readjusting in response to changes in these conditions. Thereby overall performance of TCP is degraded through additional retransmission and wasted opportunities in maintaining the communication pipe full. There have been numerous attempts to improve TCP in wireless environments, for example, Chandran, Raghunathan, Venkatesan, and Prakash (2001), Kim, Toh, and Choi (2000) and Liu and Singh (2001). A standard implementation (Wireless Profiled TCP) is defined by WAP forum, which comprises state-of-the-art works on the subject, and implementation of TCP stack in modern operation systems supports them by default (Macdonald & Barkley, 2000).

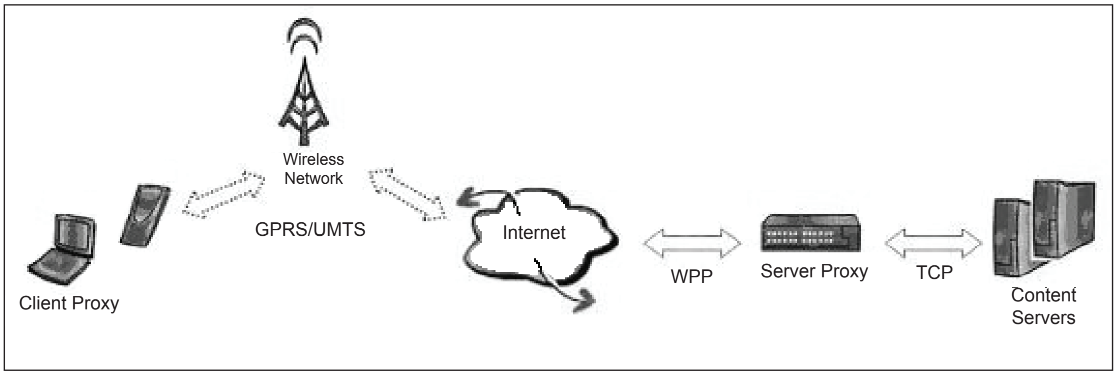
As an addition to TCP optimization, the performance enhancing proxy (PEP) concept was developed (Border, Kojo, Griner, Montenegro, & Shelby, 2001) to further improve performance of Internet applications over wireless links. Different types of PEPs are used in different environments to overcome different link characteristics, which affect the performance.

Kustov et al. (2002) proposed an idea for raising efficiency of data transfer over high-latency low-bandwidth links by combining application-level PEP and TCP replacement with a lightweight protocol stack. This approach is called protocol replacement proxy (PRP).

As shown on Figure 1, data transfer over wireless link is performed using WAP peer protocol (WPP). At both ends of the link client and server proxies perform protocol translation between TCP and WPP. For example, when a mobile user downloads a Web page from a content server, the client proxy accepts a TCP connection from the browser, passes the request to the server proxy, which, in turn, establishes a TCP connection to the content server. The content server response is delivered to the browser in the same way.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. PEP approach with transport protocol replacement



WPP stack uses UDP as a transport protocol and utilizes WTP and WDP layers of standard WAP protocol stack, with proprietary WPP layer added on top of WTP layer. This scheme makes use of both application-specific optimization techniques (e.g., for HTTP metadata, applets removal, lossy image compression), and transport protocol optimization, which is achieved by reducing overhead and adapting it to wireless link characteristics.

PROBLEM DEFINITION

The concept of PRP, proposed by Kustov et al. (2002) was experimentally studied in Kustov and Lang (2005) for its practical realization (Nokia Wireless Accelerator). The study focused on mobile Internet experience for the two most commonly used Internet protocols—FTP and HTTP—for network technologies ranging from GPRS to UMTS. Application-specific optimization depends strongly on the transferred data; for example, for office documents and HTTP browsing, data reduction was in the range of 60-90%. Transport-level optimization was found to reach 35% in the case of GPRS, which was the main target for PRP.

At the same time, it was found that the wireless accelerator yields negative time savings in the 3G network. Taking into account the fact that compression was applied to the downloaded data, all this indicated inefficient bandwidth utilization of WPP stack compared to that of wireless profiled TCP. The reason was found to be static protocol sending window, decreasing efficiency of WPP at higher data rates. Initially optimized for slower GPRS/EDGE, the same window only allowed partial bandwidth utilization in 3G.

For PRP to work properly in 3G networks as well, WPP protocol stack needs improvements.

DYNAMIC WPP STACK ADAPTATION

In this article we propose the solution to the problem—correction of the WPP protocol stack, allowing the transport layer to adapt in real-time to available bandwidth.

The optimal size of sending window is equal to bandwidth-delay product. As this value can not be measured directly, we have developed the method of adaptation based on adjusting sending window size according to bandwidth utilization feedback. For that, all packets are supplied before sending with time intervals placed in transport information items (TPI) of WTP packet headers. At the receiving side, time intervals between the moments of arrival of two consequent packets are compared to the intervals between the packets departure.

Figure 2 explains the idea in detail. Time interval **t1** between sending of the first and the second packets is conveyed from the sender to the receiver within the second packet. At the receiving side, the interval between times of arrival **t1'** is measured and then the difference (**t1'** – **t1**) is sent as feedback to the sender within WTP ACK packet.

If the difference is positive like in the case (a), this means that packets are sent faster than it is possible to deliver with currently available network bandwidth. In the long run it might result in packet losses and thus retransmissions. To avoid this, sending window size has to be decreased in order to restrict the data-sending rate.

If the difference is negative like in the case (b), the network is capable to deliver packets faster than they are currently sent, so available bandwidth is not utilized fully and sending window size has to be increased.

Figure 2. Window adaptation algorithm for WPP protocol stack improvement

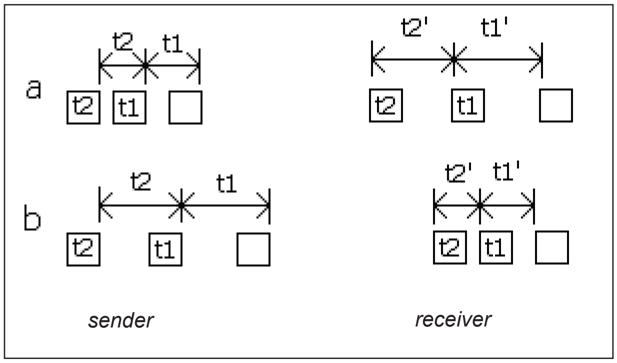
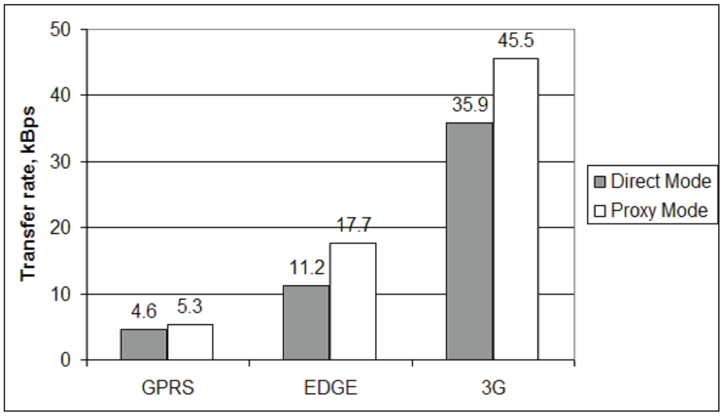


Figure 3. Improved WPP (Proxy) in comparison with TCP (Direct)



To avoid possible oscillations, the feedback values being received at the sending side are smoothed before they are used to adjust sending window size.

Predefined limits of sending window size are set to protect adaptation system from running out of reasonable boundaries.

Besides sending window size, retransmission timeout is adjusted based on current RTT value, which is measured as the time interval between sending a data packet from WTP layer and receiving corresponding acknowledgement packet (ACK) from the peer WTP layer.

EXPERIMENTAL VERIFICATION

We measured efficiency of the improved PRP scheme in GPRS/EDGE and 3G networks. We compared two file downloads: a) in direct mode, where the mobile client directly connects to the origin content server and standard TCP stack is used; and b) in proxy mode, where a tunnel is established between mobile client and PRP proxy over a wireless link and WPP protocol is used (Figure 1). As aforementioned WPP changes don't affect application-specific optimization, we focused on measuring transport protocol efficiency. To minimize the effect of compression performed by the proxy at application level, we downloaded an almost incompressible MP3 file. We have chosen large enough file (>800 KB) for its download to take time long enough to measure it even in the faster 3G network.

Measurements were done in live GPRS/EDGE/3G networks in the downtown of Helsinki. We repeated every measurement until we could reliably identify inconsistent results caused by delay spikes; after dropping them, from the remaining data we calculated average values of downlink transfer rate. Results are shown in Figure 3.

We can see that PRP with improved WPP stack provides data rates higher than TCP for all networks, improvement being at least 15% (GPRS) and reaching 60% for EDGE, with 3G results being in between (27%). The measurements in proxy mode must be corrected with accordance of data reduction, as proxy compresses all data traversing the tunnel. PRP performs gzip compression on-the-fly on chunks of the data stream, so a compression ratio that is not as high as could be achieved by applying standard archiving programs to the whole file was measured as a ratio of amount of data written to the client application (a browser) to the amount of data counted at WDP (i.e., UDP) layer. For the MP3 file that we used in measurements, PRP compression ratio was only 6%; this proves that the increase in performance gained with PRP is essentially due to more efficient transport protocol.

DISCUSSION

In this article we finalized our study of performance replacing proxy solution for efficient Internet access in mobile networks, which utilizes content optimization at application layer as well as TCP protocol replacement at transport layer with WPP, a lightweight UDP-based protocol.

Kustov and Lang (2005) pointed out bandwidth utilization inefficiency in WPP transport, resulting in low overall PRP performance in a 3G network, although content optimization at the application level is capable of reducing traffic over a wireless link up to 60-90%. This presented a problem when PRP optimization was used in latest mobile handsets, which support multiple network standards and can switch between them automatically.

To overcome the problem, in this article we proposed an algorithm for adapting of the sending window to available network bandwidth and RTT. The algorithm is based on real-time bandwidth probing from data stream, which allows extending the PRP solution applicability from slower 2.5G to faster 3G networks.

We measured the efficiency of improved WPP transport in comparison with TCP, and we have shown that WPP transport outperforms TCP in the whole range of bandwidth; the data rate increase in PRP case was in the range of 15-60%, although the data compression at PRP was only 6%.

Combined with payload reduction, better efficiency of its transportation with WPP makes PRP solution an even more efficient way of improving user experience.

Obviously needed in slower GPRS networks, optimization is also relevant even for relatively fast 3G networks; at least as long as a user is billed for traffic and network coverage restricts data rate at the levels far from theoretical maximum.

REFERENCES

Border, J., Kojo, M., Griner, J., Montenegro, G., & Shelby, Z. (2001). *RFC 3135 performance enhancing proxies intended to mitigate link-related degradations.* Internet Informational RFC 3135.

Chandran, K., Raghunathan, S., Venkatesan, S., & Prakash. R. (2001). A feedback-based scheme for improving TCP performance in ad hoc wireless networks. *IEEE Personal Communications*, *8*(1), 34-39.

Gurtov, A. (2004) *Efficient data transport in wireless overlay networks*. PhD Thesis. University of Helsinki.

Inamura, H., Montenegro, G., Ludwig, R., Gurtov, A., & Khafizov, F. (Eds.). (2003). *RFC 3481 TCP over second (2.5G) and third (3G) generation wireless networks*.

Kim, D., Toh, C.-K., & Choi, Y. (2000). TCP-BuS: Improving TCP performance in wireless ad hoc networks. In *IEEE International Conference on Communications, ICC 2000* (pp. 1707-1713).

Kustov, A., Auvinen, O., Hämäläinen, M., Kari, H., Khachtchanski, V., Koponen, J., Mallat, H., Räsänen, J. (2002). *Methods and arrangements for providing efficient information transfer over a limited speed communications link*. US Patent # US2002138565.

Kustov, A., & Lang, J. (2005). Protocol replacement proxy for improving efficiency of mobile Internet access. In *Proceedings of the Third International Conference on Advances in Mobile Multimedia* (pp. 355-364). Austrian Computer Society.

Liu, J., & Singh, S. (2001). ATCP: TCP for mobile ad hoc networks. *IEEE Journal on Selected Areas in Communications*, *19*(7), 1300-1315.

Macdonald, D., & Barkley, W. (2000). *Microsoft Windows 2000 TCP/IP implementation details*. Retrieved from http://www.microsoft.com/technet/itsolutions/network/deploy/depovg/tcpip2k.mspx

Nokia Wireless Accelerator. (n.d.). Retrieved from http://europe.nokia.com/nokia/0,0,77182,0.html

Tsaoussidis, V., & Matta, I. (2002). Open issues on TCP for mobile computing. *The Journal of Wireless Communications and Mobile Computing*, *2*(1), 3-20.

Wireless Profiled TCP. (2001). Retrieved from http://www.openmobilealliance.org/release_program/docs/Browser_Protocol_Stack/V2_1-20050204/WAP-225-TCP-20010331-a.pdf

KEY TERMS

Bandwidth Utilization: A criterion to measure the ability of a system to efficiently use radio data channel; it reaches its maximum value when data fill the channel as much as possible.

Content Optimization: Minimizing data redundancy by means of, for example, applying compression in order to reduce traffic over wireless channel.

Bandwidth Probing: A method to measure link capacity by sending dedicated test packets, or appending service data to payload.

Sending Window: At the transport protocol layer, amount of data in-the-flight; that is, packets that have been sent but have not been acknowledged by the receiver.

Retransmission Timeout (RTO): It is a time interval for the sender to wait for the packet acknowledgement from the receiver before starting retransmission.

Round Trip Time (RTT): It is a measure of the time it takes for a packet to travel from a mobile terminal, across a mobile network to a server, and back.

Performance Enhancing Proxy (PEP): A technique employed to improve degraded TCP performance caused by characteristics of specific link, for example, in wireless environments.

Providing Location-Based Services under Web Services Framework

Jihong Guan
Tongji University, China

Shuigeng Zhou
Fudan University, China

Jiaogen Zhou
Wuhan University, China

Fubao Zhu
Wuhan University, China

INTRODUCTION

Location-based services (LBSs) provide personalized services to the subscriber based on his or her current position. By combining information on the location of a mobile user, services can be tailored exactly to the user's situation. These powerful location-based services are the keys to the growth of the mobile Internet in both the consumer and business markets. The geographical location of mobile phones can be used by operators and service providers to create and offer location-based mobile services, which employ accurate, real-time positioning to connect users to nearby points of interest, advise them of current conditions such as traffic and weather, or provide routing and tracking information—all via wireless devices. For subscribers, this means access to attractive, convenient, value-added services that will make their lives easier and more fun, save time, enhance business efficiency, and increase personal safety (Anckar & D'Incau, 2002).

To meet the demands of efficient, stable, and scalable architecture and implementation techniques of location-based mobile services, adjusting for various multi-platforms, multiple applications, and sustainable development environment, we propose an architecture of LBS based on Web services technologies—that is, Web service-based LBS (WS-LBS). WS-LBS has multi-layers and consists of a database server, global spatial information servers, local spatial information servers, and mobile clients. Web services technologies are adopted in WS-LBS. UDDI is used to publish global spatial information services, and URL address is used for local spatial information services publishing. At the mobile client site, SOAP client technology is adopted for end users to access remote spatial information services. Two kinds of WS-LBS clients are implemented: J2ME client and WinCE client. A WS-LBS prototype is implemented by Java and c# languages, providing transparent access to distributed spatial information services for various mobile end users.

BACKGROUND

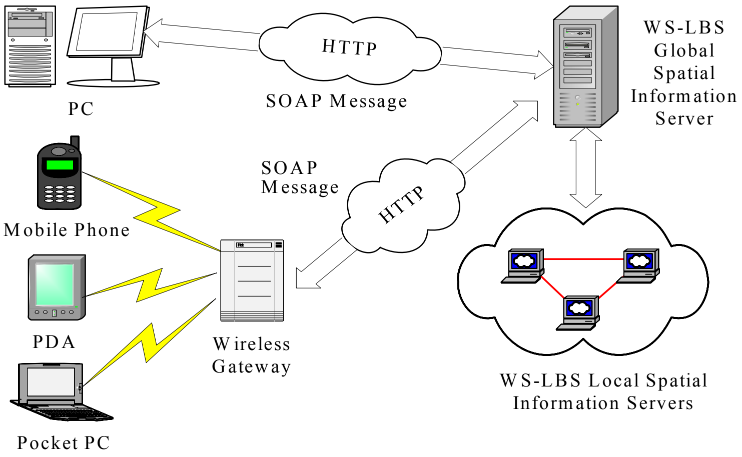
Mobile terminals such as cellular phones, PDAs, palmtops, and so forth emerge as a new class of small-scale, ad-hoc service providers and consumers. And location-based services have been a rapidly growing concept in telecommunication industry. Market research companies predict a huge market for services to be delivered to mobile users. Strategy Analytics, a leader in providing strategic and tactical support for business planners, recently concluded that: "Demand for mobile information services is skyrocketing and interest in coupling them with positioning technologies [is] at an all time high" (Raskind, 2000). Location technologies are expected to augment existing wireless applications as well as spawn a host of entirely new services, including alerts, advertisements, and personal location and guidance services. LBSs are services that are triggered by the current geographic location of the mobile user and his surroundings. According to ARC Group Consultants (2003), it is expected that LBSs will be the most widely used mobile services by 2007.

LBSs can be categorized into three main classes (Beinat, 2001):

- **Information Services:** Providing information about objects close to the user—services like "Where is an ATM nearby" or "Find the nearest parking lot."
- **Interaction Services:** Based on the interaction between mobile users/objects and which do not require a "mobile Internet" component or content sources.
- **Mobility Services:** Supporting smart mobility and revolving around navigation capabilities. An example may be: "How to get to the nearest hospital?"

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. An overview of the WS-LBS architecture


HTTP
SOAP Message
PC
WS-LBS Global Spatial Information Server
SOAP Message
HTTP
Mobile Phone
PDA
Wireless Gateway
Pocket PC
WS-LBS Local Spatial Information Servers

Location-based services provide personalized services to the subscriber based on his or her current position (Searby, 2003). By combining information on the location of a mobile user, services can be tailored exactly to the user's situation. These powerful location-based services are the keys to the growth of the mobile Internet in both the consumer and business markets. For subscribers, this means access to attractive, convenient, value-added services that will make their lives easier and more fun, save time, enhance business efficiency, and increase personal safety.

Many GIS enterprises bring forth their wireless schemes: ESRI ArcPad is integrated with GPS module to provide mobile users functions similar to desktop GISs; ArcLocation gives a set of wireless solutions; and MapInfo's miAware, Intergraph's IntelliWhere, and Oracle's Mobile Location Services on Oracle8 and Oracle 10g are all conformed with OpenLS specifications of OGC. Mobile terminal providers such as Nokia, Ericsson, and Motorola are also competitive to develop added-value businesses for mobile users. However, the architecture, platforms, standards, and application background of these systems and applications are quite different (Adams, Ashwell, & Baxter, 2003). To keep sustainable development of LBSs, urgent need exists for efficient, stable, and scalable architecture and implementation techniques of location-based mobile services, adjusting for various multi-platforms, multiple applications, and sustainable development environment (Hermann & Heidmann, 2002).

Web services as a language-neutral and platform-independent technology can be adopted in constructing flexible and loosely coupled business systems. It is easy to apply Web services as wrapping technology around existing applications and information technology assets, and new solutions can be deployed quickly and recomposed to address new opportunities under a Web services framework (Gottschalk, Graham, Kreger, & Snell, 2002). Hence, we propose a Web services-based LBS architecture: Web service-based LBS (WS-LBS).

WS-LBS ARCHITECTURE

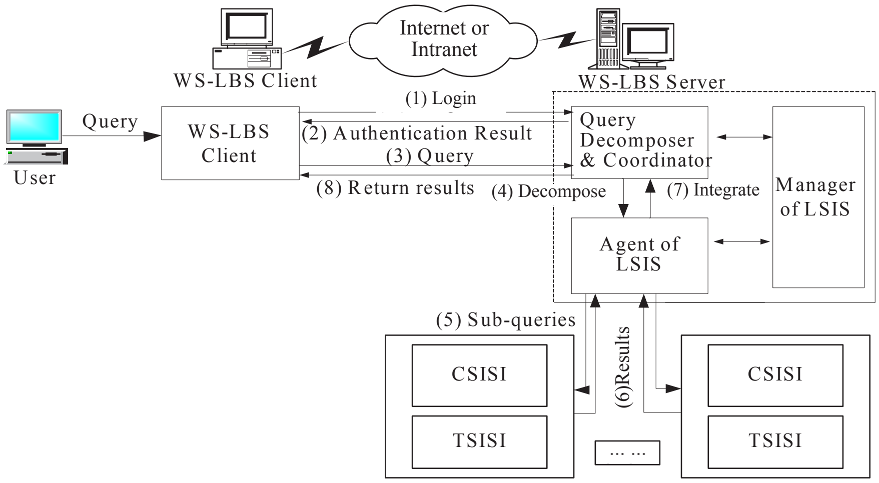
The WS-LBS system consists of four components: client site, server site, Internet or intranet connecting the client sites and server sites, and SOAP messages among the sites. Figure 1 is an overview of the WS-LBS architecture.

In the WS-LBS system, a client is referred to a client machine that can be a desktop or a notebook computer, or a PDA or a mobile phone, which usually does not provide functions of spatial information processing. A user can submit a query request and obtain the final results via a client machine, and only if the user wants to connect to WS-LBS to get spatial information needed, then the client machine can be a part of system. A server in WS-LBS usually refers to a spatial information server that can provide spatial services for local and remote users. When a user submits a query via a client, the query is encapsulated in a SOAP message after initial analysis, and the message is sent to a related server. After receiving this SOAP message, the server makes further analysis and optimization on the requirement, and decomposes the query into sub-queries. Then one or multiple SOAP messages will be sent to related server(s) for the required information or services. The search results will return to the original server within SOAP messages and be merged into final results back to the client.

The Framework of the WS-LBS Server

The framework of the WS-LBS server has two layers: the global spatial information services (GSIS) layer and the local spatial information service (LSIS) layer. GSIS consists of three components: agent of LSIS, manager of LSIS, and query decomposer and coordinator. LSIS is composed of two components: common spatial information services interfaces (CSISI) and thematic spatial information services interfaces (TSISI).

Figure 2. Query handling procedure in WS-LBS



Agent of LSIS in GSIS provides direct proxy for those services with independent functions and no necessity of cooperation with other servers. With the proxy, terminal users need not know the exact providers of such spatial information services. Manager of LSIS is responsible for the management of LSIS, such as registration of new services, or destroy, update, suspend, or activate services, and so forth. The manager also has duties of statistically analysis, tracing, and monitoring the running status of LSIS to improve the holistic quality of services. Query decomposer and coordinator takes charge of the decomposition of the user's query into sub-queries, and the coordination of such sub-queries is sent to corresponding services providers, along with the merge or integration of the searched results so as to return the final result to the user.

CSISI is mainly composed of authentication interface, local metadata interface, map generation interface, non-spatial query interfaces, and so forth. TSISI consists of certain spatial operation service interfaces, such as interfaces of adjacent relationship query, real-time tracing, coverage analysis, route analysis, and so forth.

WS-LBS Client Site

A WS-LBS client is a computer or a mobile device that a user used to access the geographic information services in WS-LBS. Figure 2 shows the request handling procedure of a user's query. At the client site, a user logs into a WS-LBS server and waits for authentication; he or she then submits his or her request to the server. The server will make further analysis, decomposition, and optimization on the query, and take charge of finding, coordinating, merging, and returning the results to the user.

Since WS-LBS is based on Web services technologies to construct and publish geographic information services, at the client site we only need to adopt SOAP client technology to access remote geographic information services.

WS-BASED SPATIAL SERVICES ESTABLISHING AND PUBLISHING

The WS-LBS server mainly aims to effectively provide various spatial information services. For services creation, two approaches can be applied: services defined with WSDL or defined with Java Interfaces description. See examples 1 and 2.

Services can be classified into global or local ones. Considering services publishing in WS-LBS, UDDI is adopted to publish global geographic information services, which

Example 1.

```
// Java Interfaces definition to describe spatial information services
package CAN ;
    public class MapService
    {   // transform map layer to a SVG document
        /* @jws:operation*/
        public String Layer2Svg(String strLayer){}
        // transform map layer to a GML document
        /* @jws:operation*/
        public String Layer2Gml(String strLayer){}
        // integrate map layers into a SVG document
        /* @jws:operation*/
        public String IntegrateSvg(String strLayers){}
    …
    }
```

Example 2.

```
// WSDL based description of spatial information services:
<?xml version="1.0" encoding="utf-8"?>
< definitions xmlns="http://schemas.xmlsoap.org/wsdl/"
    xmlns:conv="http://www.openuri.org/2002/04/soap/conversation/"
    xmlns:cw="http://www.openuri.org/2002/04/wsdl/conversation/"
    xmlns:http="http://schemas.xmlsoap.org/wsdl/http/"
    xmlns:jms="http://www.openuri.org/2002/04/wsdl/jms/"
    xmlns:mime="http://schemas.xmlsoap.org/wsdl/mime/"
    xmlns:s="http://www.w3.org/2001/XMLSchema"
    xmlns:s0="http://www.openuri.org/"
    xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/"
    xmlns:soapenc="http://schemas.xmlsoap.org/soap/encoding/"
    xmlns:xm="http://www.bea.com/2002/04/xmlmap/" targetNamespace="http://www.openuri.org/" >
    <types>
        <s:schema   attributeFormDefault="qualified"   elementFormDefault="qualified"   targetNamespace="http://www.openuri.org/">
        <s:element name="Layer2Svg">
                <s:complexType>
                    <s:sequence>
                    <s:element minOccurs="0" maxOccurs="1" name="strLayer" type="s:string" />
                </s:sequence>
                </s:complexType>
        </s:element>
        ……
    </s:schema>
</types>
<message name="Layer2SvgSoapIn">
    <part name="parameters" element="s0:Layer2Svg" />
</message>
……
<portType name="MapServiceSoap">
    <operation name="Layer2Svg">
        <input message="s0:Layer2SvgSoapIn" />
        <output message="s0:Layer2SvgSoapOut" />
    </operation>
    ……
</portType>
    …...
<binding name="MapServiceSoap" type="s0:MapServiceSoap">
    <soap:binding transport="http://schemas.xmlsoap.org/soap/http" style="document" />
    <operation name="Layer2Svg">
        <soap:operation soapAction="http://www.openuri.org/Layer2Svg" style="document" />
        <input soap:body use="literal" />
        <output soap:body use="literal" />
    </operation>
        ……
</binding>
    ……
<service name="MapService">
    <port name="MapServiceSoap" binding="s0:MapServiceSoap">
        <soap:address location="http://192.168.0.35:7001/samples/CAN/MapService.jws" />
    </port>
    ……
    </service>
</definitions>
```

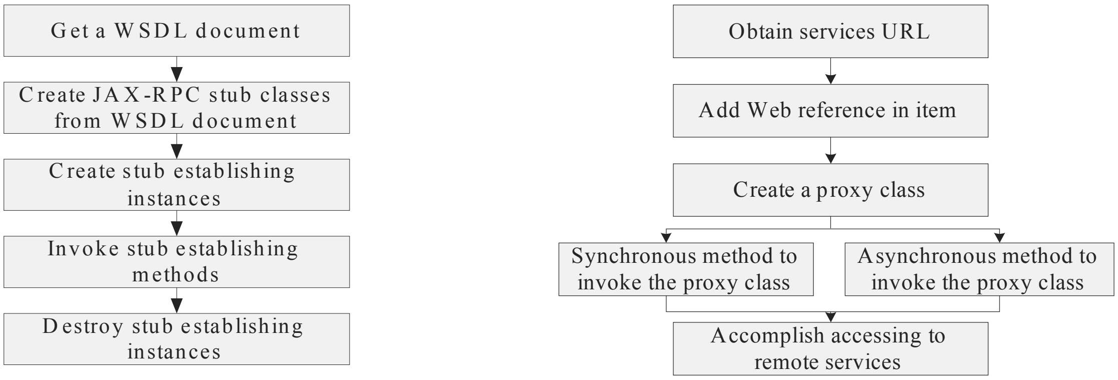
have relatively stable interfaces, while URL address is used for local geographic information services publishing, which have more flexible interfaces.

SPATIAL SERVICES ACCESS VIA MOBILE DEVICES

For mobile devices such as cellular phones, PDAs, palmtops, and so forth, embedded operation systems, application platforms, and software are used. For example, Symbian, Windows CE, Palm OS, and Linux are such embedded OS, while BREW, J2ME, and .net are the software used often in mobile terminals.

Windows CE is Windows Consumer Electronics or simply WinCE. It is Microsoft's version of Windows for handheld devices and embedded systems that use x86, ARM, MIPS, and SHx CPUs. J2ME stands for Java 2 Platform, Micro Edition. It is a version of Java 2 for cell phones, PDAs, and consumer appliances. J2ME uses the K

Figure 3. Web services accessing under J2ME and WinCE environment


Get a WSDL document
Create JAX-RPC stub classes from WSDL document
Create stub establishing instances
Invoke stub establishing methods
Destroy stub establishing instances
Obtain services URL
Add Web reference in item
Create a proxy class
Synchronous method to invoke the proxy class
Asynchronous method to invoke the proxy class
Accomplish accessing to remote services

(a) accessing based or J2ME

(b) accessing based on WinCE

Virtual Machine (KVM), a specialized Java interpreter for devices with limited memory. The connected limited device configuration (CLDC) provides the programming interface for wireless applications. The mobile information device profile (MIDP) provides support for a graphical interface, networking, and storage.

There are three main schemes (WAP, J2ME, or WinCE based) for accessing services via mobile terminals. A WAP-based solution consists of a WAP gateway, WAP mobile phones, and spatial information servers. WAP gateway is applied to bridge the World Wide Web and a mobile network like a protocol translator. The channel width, storage, and process ability that the WAP solution depends on are limited to provide effective multimedia services. In WS-LBS, we design and implement the other two schemes: J2ME- and WinCE-based schemes.

Scheme Based on J2ME Technology

Mobile terminal accessing scheme based on J2ME is neutral to platform, which can provide support for PDAs, cellular phones, TV set top boxes, remote controllers, and other embedded devices. There are WS-APIs provided in J2ME. The client site implementation procedure is shown in Figure 3(a). First, get a WSDL document, next create JAX-RPC stub classes, then through the stub classes the client can access Web services.

Scheme Based on WinCE Technology

WinCE is developed specially by Microsoft to support handsets and information household appliances. A services accessing procedure implemented on WinCE includes these steps: first, obtain the services URLs, and add Web reference in the item, create a proxy class, and use synchronous or asynchronous accessing method to invoke the proxy class; finally accomplish the access to remote services. The procedure is shown in Figure 3(b). The synchronous method is simple, while the asynchronous one is relatively complex. If there is no response from the Web services for a long time, synchronous invoke will freeze the application, while asynchronous invoke allows users to interact with application during invoke procedure.

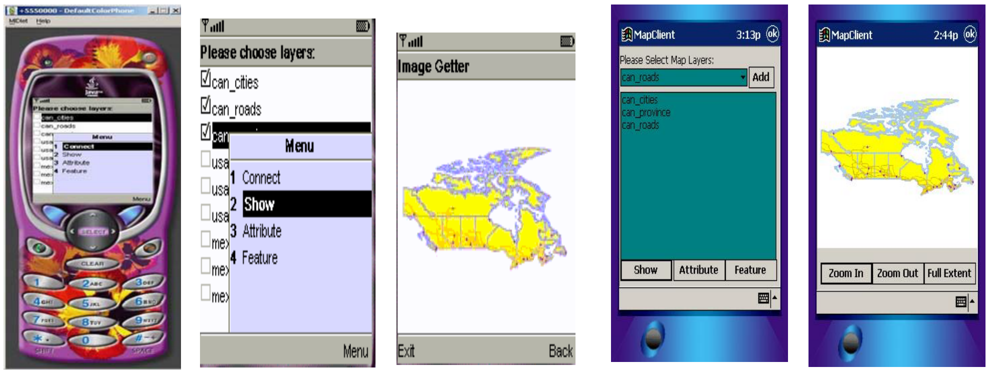
WS-LBS PROTOTYPE

WS-LBS prototype is implemented by Java and c#. The development tools include Weblogic Platform 7, J2ME Wireless Toolkit 2.1, Microsoft Visual Studio 2003, Oracle9i and Oracle9i Spatial, JDK1.4.1, DOM4J, Geotools0.8.0, Batik1.5.1, SQL Server 2000, and Jbuilder. Five desktop computers and one portable computer connected via Internet are designed as servers and clients in WS-LBS, and two kinds of mobile phone simulators are included in the system. The geographic information and some spatial information services are distributed to machines among the WS-LBS system. Users can get some simplified location-based services via mobile phone simulators. The interfaces and menu selection example and a query sample are shown in Figure 4.

CONCLUSION

LBS is an value-added business which gets a user's location information through a telecom provider's network and provides the corresponding services under the support of a digital map platform. In this article, we propose to provide LBS under a Web services framework. We call such services: Web services-based LBS, or simply WS-LBS. The architecture and implementation techniques of WS-LBS are given, and a WS-LBS prototype is developed, which

Figure 4. Query interface and examples at J2ME and WinCE client sites respectively



(a) J2ME based Cl,ient

(b) WinCE based Client

provides transparent access to distributed spatial information services via mobile terminals.

Research and applications of LBS involve not only spatial information technologies, but also mobile telecom technologies. Furthermore, technologies of wireless markup language, mobile database, mobile localization, mobile network, mobile terminals, wireless application server, mobile connection, as well as appropriate architecture, development platform and software, and efficient data management for LBS are all under deep investigation.

ACKNOWLEDGMENTS

This work was supported by grants numbered 60573183 and 60373019 from NSFC, grant No. 20045006071-16 from Chenguang Program of Wuhan Municipality, grant No. WKL(04)0303 from the Open Researches Fund Program of LIESMARS, and the Shuguang Scholar Program of Shanghai Education Development Foundation.

REFERENCES

Adams, P., Ashwell, G., & Baxter, R. (2003). Location-based services—An overview of the standards. *BT Technology Journal, 21*(1), 34-43.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *The Journal of Information Technology Theory and Application, 4*(1), 43-64.

ARC Group. (2003). Retrieved from http://www.arcgroup.com/index.html

Beinat, E. (2001, April). Location-based services, market and business drivers. *Geoinformatics Magazine.*

ERSI. (n.d.). Retrieved from http://www.esri.com/

Gottschalk, K., Graham, S., Kreger, H., & Snell, J. (2002). Introduction to Web services architecture. *IBM Systems Journal, 41*(2), 170-177.

Harry, N., & Image, M. (2002). *OpenLS architecture & application overview.* Retrieved from http://www.openls.org/dvdl/tsl/

Hermann, F., & Heidmann, F. (2002). User requirement analysis and interface conception for a mobile, location-based fair guide. *Proceedings of the 4th International Symposium on Human Computer Interaction with Mobile Devices* (LNCS 2411, pp. 388-392).

Intergraph Corporation. (n.d.). Retrieved from http://www.intergraph.com/

Kurt, B. (2002). *OGC and LBS overview.* Retrieved from http://www.openls.org/dvdl/tsl/

MapInfo Corporation. (n.d.). Retrieved from http://www.mapinfo.com/industries/mobile/solution_lbs_platform.cfm

Raskind, C. (2000). *Location-based services: Revenues & applications.* Retrieved from http://www.strategyanalytics.net/default.aspx?mod=ReportAbstractViewer&a0=608

OGC. (n.d.). *OpenLS.* Retrieved from http://xml.coverpages.org/OpenLS-RFC.html

Oracle Corporation. (n.d.). Retrieved from http://www.oracle.com/

Searby, S. (2003). Personalisation—An overview of its use and potential. *BT Technology Journal, 21*(1), 13-19.

KEY TERMS

Information Services: One of the three main classes of LBSs, providing information about objects close to the user; includes services like "Where is an ATM nearby" or "Find the nearest parking lot."

Interaction Services: One of the three main classes of LBSs, based on the interaction between mobile users/objects and does not require a "mobile Internet" component or content sources.

Java 2 Platform, Micro Edition (J2ME): A version of Java 2 for cell phones, PDAs, and consumer appliances. J2ME uses the K Virtual Machine (KVM), a specialized Java interpreter for devices with limited memory. The Connected Limited Device Configuration (CLDC) provides the programming interface for wireless applications. The Mobile Information Device Profile (MIDP) provides support for a graphical interface, networking, and storage.

Location-Based Service (LBS): Used to provide personalized services to the subscriber based on his or her current position. By combining information on the location of a mobile user, services can be tailored exactly to the user's situation.

Mobility Services: one of the three main classes of LBSs, supporting smart mobility and revolving around navigation capabilities. An example may be: "How to get to the nearest hospital?"

Mobile Terminal: Examples include cellular phones, PDAs, palmtops, and so forth; emerging as a new class of small-scale, ad-hoc service providers and consumers in Internet applications, usually connected to the network via wireless connection.

Simple Object Access Protocol (SOAP): A message-based protocol based on XML for accessing services on the Web. Initiated by Microsoft, IBM, and others, it employs XML syntax to send text commands across the Internet using HTTP. SOAP is similar in purpose to the DCOM and CORBA distributed object systems, but is lighter in weight and less programming intensive.

Web Service: Can be seen as an interface that describes a collection of operations that are network accessible through standardized XML messaging. Software applications written in various programming languages and running on various platforms can use Web services to exchange data over computer networks due to the interoperability using of open standards.

Windows Consumer Electronics or Windows CE (WinCE): Microsoft's version of Windows for handheld devices and embedded systems that use x86, ARM, MIPS, and SHx CPUs.

Web Services Description Language (WSDL): An XML language for describing Web services. This is an XML-based service.

Provisioning of Multimedia Applications across Heterogeneous All-IP Networks

Michail Tsagkaropoulos
University of Patras, Greece

Ilias Politis
University of Patras, Greece

Tasos Dagiuklas
Technical Institute of Messolonghi, Greece

Stavros Kotsopoulos
University of Patras, Greece

INTRODUCTION

With the opening of the telecommunication market and the emergence of low-cost and heterogeneous wireless access technologies, it is envisaged that next-generation network and service providers will not only vary in the deployed access technology but also in their business models and structures. Such providers will differ from large providers such as the current telecom providers offering multiple services and covering large geographical areas, down to small providers offering certain services such as conferencing or messaging only or covering small geographical areas such as a coffee shop or a shopping mall. Further, while in the current networking environment, a home provider of a user is usually represented by a large telecom provider; in such a heterogeneous environment, any trustworthy entity such as an application provider, a banking entity, or a credit card provider that is capable of authenticating the user and marinating his usage profile can act as a home provider. Towards this vision this article discusses the issues that concern the establishment of multimedia applications across heterogeneous networks.

NEXT-GENERATION NETWORKS AND THE ALL-IP CONVERGENCE

Convergence of heterogeneous wireless technologies over a broadband IP core network will allow mobile subscribers to access a new variety of services, over a variety of access networks and by using a variety of devices. This integration will be realized on the network access with devices able to hand off across heterogeneous wireless access technologies, service delivery, and availability (Dagiuklas & Velentzas, 2003). There is no industry consensus on what next-generation networks will look like but, as far as the next-generation networks are concerned, ideas and concepts include:

- transition to an “All-IP” network infrastructure;
- support of heterogeneous access technologies (e.g., UTRAN, WLANs, WiMAX, xDSL, etc.);
- VoIP substitution of the pure voice circuit switching;
- seamless handovers across both homogeneous and heterogeneous wireless technologies;
- mobility, nomadicity, and QoS support on or above the IP layer;
- provisioning of triple-play services creating a service bundle of unifying video, voice, and Internet;
- home networks opening new doors to the telecommunication sector and network providers;
- unified control architecture to manage application and services; and
- convergence among network and services.

HETEROGENEOUS MULTIMEDIA NETWORKS AND SERVICES

Vision

In order to allow seamless communication and roaming in a heterogeneous wireless environment, one needs to provide an efficient way for coupling multimedia service provisioning, access with fast-handover schemes, and establishing trust relations among service providers and users. This necessitates the provisioning of a framework for establishing security and trust relations among network operators, service providers, and mobile users, allowing thereby smooth roaming among different administrative domains/networks

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Inter-domain framework

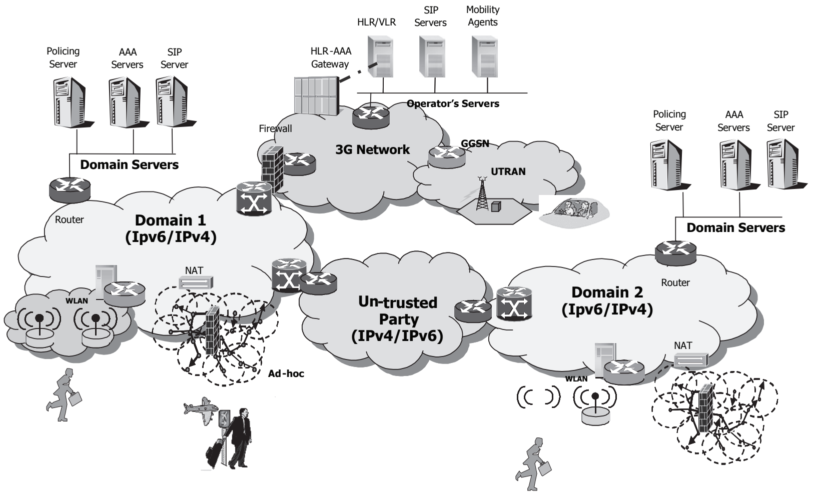
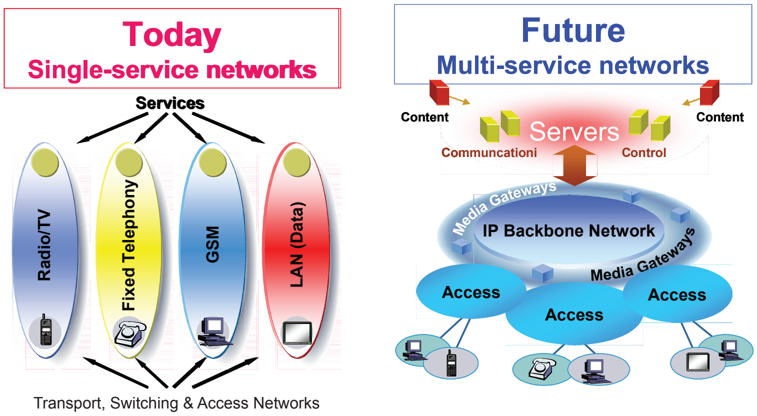

HLR/VLR
SIP Servers
Mobility Agents
Policing Server
AAA Servers
SIP Server
HLR-AAA Gateway
Operator's Servers
Firewall
3G Network
GGSN
UTRAN
Domain Servers
Router
Domain 1 (Ipv6/IPv4)
NAT
WLAN
Ad-hoc
Un-trusted Party (IPv4/IPv6)
Policing Server
AAA Servers
SIP Server
Domain Servers
Router
Domain 2 (Ipv6/IPv4)
NAT
WLAN

Figure 2. NGN architecture migration (Nokia, 2001)


Today
Single-service networks
Services
Radio/TV
Fixed Telephony
GSM
LAN (Data)
Transport, Switching & Access Networks
Future
Multi-service networks
Content
Content
Servers
Communcationi
Control
Media Gateways
IP Backbone Network
Media Gateways
Access
Access
Access

and seamless provisioning of multimedia services. Such a vision infrastructure is illustrated in Figure 1.

This infrastructure will support the dynamic establishment of trust relations between independent providers (e.g., foreign and home providers) in a distributed manner over hybrid IPv4 and IPv6 networks (Salkintzis, 2004). Moreover, it will provide the required enhancements for providing secure interconnection among different heterogeneous networks, establishing user-provider trust relations, and the necessary means for authenticating users in foreign domains and exchanging their profiles in a secure manner. This would thereby enable users to roam to foreign networks and use the provided services in these networks without affecting their privacy. Finally, to support the smooth and fast handover, efficient and secure context exchange mechanisms will be provided, allowing users to roam among different providers without having to explicitly re-authenticate themselves and establish new trust relations.

It is envisioned that the NGN architecture will be based on packet-based technologies. The most important part of NGN is the division of network functionality into many distributed functions, which fall into the following categories (Dagiuklas et al., 2005):

1. Control, management, and signaling, which provides the intelligence needed for user control of the connection. This intelligence is distributed.
2. Access, routing, switching, and transport, which provides the functions needed for transporting information between end users and other network elements.
3. Convergence with existing legacy networks (PSTN, SS7, mobile networks).

Figure 2 illustrates the transitions from the current scene towards NGN.

To achieve the goal of enabling flexible roaming between different network and services providers in a secure and chargeable manner between providers of different technologies and business models, this article will be addressing the following requirements (Velentzas & Dagiuklas, 2005):

- secure and scalable inter-domain AAA infrastructure for mobile user (secure roaming);
- user access security;
- application level security; and
- concepts for security context transfer while the user hands over among heterogeneous wireless technologies.

INTER-PROVIDER RELATIONSHIP

The establishment of heterogeneous networks aims to import substantial services for the customer through the specification and realizing of infrastructure that enable dynamic trust relation among different providers in a secure and scalable manner. Particular emphasis should be given to roaming business models while the mobile users move across different administrative domains. The possibly very large number of service and network providers as well as home providers makes the establishment of static trust relations between all of those providers tedious and non-economical. Therefore, mechanisms and solutions must be defined that will allow users to roam from one provider (either network operator or service provider or a mixture of both) to the other without having to explicitly subscribe to the services of each provider or assuming some pre-established trust relation between the different providers.

Dynamic Trust Relations

Most current solutions for enabling roaming of users among different network providers assume the existence of pre-established trust relations between the involved partners. Two providers that would like to enable user roaming must be registered at this broker. Currently, the number of network providers is relatively small, with those providers having a large scale in terms of the number of supported users and covered geographical area. Further, the definition of a home provider might change as well, allowing for example a banking unit to act as the home provider of a user and authenticate his identity. Registering all possible providers at a single trust broker would surely not scale. This would then lead to having different brokers each maintaining trust relations to only a small number of providers. To allow for roaming between any two entities, the foreign network needs to be able to discover the home provider of the user and which trust brokers could be used for establishing a trust relation with the home provider. The experience gained for private key distribution architectures shows that for such schemes to work, they need to be distributed. The work to be done here would include specifying a distributed trust infrastructure and investigating its applicability and performance with the increasing number of providers in NGN.

Identity Management

Identity management is best defined as “those IT and business processes, organizations, and technologies that are applied to ensure the integrity and privacy of identity and how it translates to access.” This results in its effective use as a crucial element of IT security infrastructure. Recently, identity management has been elevated within many IT organizations to be a formal program consideration by several business drivers.

The Liberty framework, developed by the Liberty Alliance project (www.projectliberty.org), comprises an identity management architecture focused on the establishment of secure, consistent, and manageable business relationships and interactions over the Internet. This architecture is realized along circles of trusts embracing affiliated identity providers and service providers. The Liberty framework further defines adequate schemas and metadata for identity management, along with the required bindings to Web-based middleware (SOAP) services.

Current roaming solutions require close interaction between the home and foreign providers in order to gain access to user identity profiles. 3GPP solutions necessitate for example that any signaling messages sent by the user be routed through the home network. This not only increases the set-up delay for establishment of communication sessions, but also complicates the offering of local services in foreign networks. To overcome this problem the home and foreign networks need to be able to exchange user profiles that allow the foreign network to make the decision of whether or not to grant a user access to certain services locally. However, while doing this the privacy of the user’s data and profile need to be guaranteed, and the foreign provider should only be given information that is needed for authorizing the user and providing him with the necessary access rights.

Policy-Driven, Configurable, and Programmable AAA Infrastructures

The exact behavior of AAA infrastructure might vary considerably depending on its location (intra- or inter-domain), functionality (broker, proxy, home, or foreign), supported features (QoS, mobility, multimedia), provider policy, load balancing architectures, or security requirements for example. To ease the replacement of RADIUS protocol-based servers to diameter-based AAA servers, server and network providers need to have powerful and yet simple interfaces for programming and customizing those servers (Nakhjiri & Nakhjiri, 2005). While Diameter-based AAA technologies are increasingly being considered as the basis for AAA in NGN, there will always be providers relying on proprietary or older solutions. To still be able to communicate with such providers and exchange AAA information with them, some translation mechanisms need to be used to cover the gap.

Reliable and Secure Intra- and Inter-Provider AAA Infrastructure

While attacking a certain server of access router by generating a lot of useless traffic might render part of the network or a certain service useless, attacking the AAA infrastructure would render the complete network useless. Mounting an attack on a trust broker would make any roaming between the networks impossible. Currently, the AAA infrastructure in PSTN networks is not very vulnerable simply because the end systems that are allowed to connect to the network are "dumb." However, in the open environment of NGN mounting attacks on an AAA server will become easier. Any system can start a large number of authentication requests and occupy thereby not only a substantial share of the network bandwidth but also of the processing resources available to the AAA servers.

Another aspect to be considered here is the reliability of the AAA infrastructure in the face of various failure situations such as software or hardware failures of the AAA servers themselves or the links connecting those servers to the networks. The following points need to be dealt with in future work:

- identification of attack and exploitation possibilities on AAA infrastructures;
- dynamic detection of attacks on AAA infrastructure and devising of mechanisms for defending and protecting against those attacks and reducing their effects; and
- specification and realization of fail-over mechanisms for AAA servers.

Secure Multimedia Service Access

In the literature of node cooperation enforcement, the proposed solutions can be subdivided into two main categories: trade-based schemes and reputation-based schemes. In trade-based schemes, a node that provides some service to a peer node (e.g., packet forwarding) is rewarded by either another immediate service in exchange or some monetary token that he can later use to buy services from another node. In reputation-based schemes each node keeps a reputation metric for other nodes it deals with and provides services only to nodes that exhibit good reputation.

In all reputation-based mechanisms for cooperation enforcement, each node in the network performs two distinct functions: rating the behavior of neighboring nodes and using these ratings to adjust its own behavior towards them. Rating the conformance of neighboring nodes to a given network protocol is an operation that depends on the specific protocol and network architecture. For instance, in single-channel MANETs, rating the packet forwarding service provided by a node's neighbors is simply performed through monitoring of the common channel. However, in clustered MANETs, which use different channels in each cluster and bridge nodes to relay packets between clusters (such as Bluetooth scatter nets), a node cannot receive the transmissions of all of his neighbors. Hence, a different technique for rating the forwarding services provided by them is needed. Similarly, rating the conformance to a neighborhood discovery protocol or a Medium Access protocol is fundamentally different than rating packet forwarding.

On the other hand, a cooperation reinforcing reputation mechanism can be easily adapted to use such behavior ratings independently of the rated service. A crucial task for this mechanism is to distinguish between perceived and actual non-cooperative behavior. For example an MT might receive a bad cooperation rating because of link failure or mobility. Misbehaving MTs might also choose to misbehave in a probabilistic way in order to evade detection. If erroneously perceived misbehavior is permitted with a certain probability, then the detection of intentional misbehavior is reduced to an estimation problem.

Secure Service Discovery

The service discovery may be performed using a hierarchical multi-tier approach based on several tiers. A service manager (SM) discovers and manages the services in its corresponding tier and interacts with its upper-tier SM. All the available public and private services provided by foreign nodes, directly attached to the network, or provided by any other kind of infrastructure network will be discovered, subject to authentication and authorization.

In the service discovery architecture, the protocols used for communication are some common service discovery protocols, such as UPnP, Bluetooth SDP, and JXTA. At the highest level, the objective is to create an environment where message-level transactions and business processes can be conducted securely in an end-to-end fashion. There is therefore a need to ensure that messages are secured during transit, with or without the presence of intermediate nodes. There may also be a need to ensure the security of the data in storage. The requirements for providing end-to-end security for the service discovery are summarized as follows:

- **Secure Service Registration and Deregistration:** Service registration and deregistration corresponds to service information management by the SMs. Only authorized service providers should be allowed to register and deregister a service from the repository. Meanwhile, it is important to maintain the integrity and confidentiality of the registered services in the service registration and deregistration process. Other types of attacks should be mitigated such as message replay and spoofing. Efficient message authentications through signatures or keyed hash-functions are suitable countermeasures for these attacks.

- **Secure Authorization:** Authorized service discovery is needed to control the discoverable services by each entity. On the one hand, users may be authorized to perform a service discovery, but based on their set of credentials, they may only have a controlled visibility of available services. On the other hand, the discovered service must be genuine and trustworthy. Finally, the anonymity of the entity performing the service discovery and confidentiality of the process should also be ensured.
- **Secure Service Delivery/Provisioning:** After a service has been found, it should be securely delivered by the mutual authentication between the recipient of the service and the service provider. Delivery confidentiality and service integrity must also be met so that the genuine and authentic service is only delivered to the intended recipient(s). As for secure service discovery, anonymity with reference to the location and identity can also be required.
- **Dependability:** Dependability can be defined as the property of a system, which always honors any legitimate requests by authorized entities. It is violated when an attacker succeeds in denying service access to legitimate users (e.g., by exhausting all the available resources through the DoS attack).
- **Service Access Control:** Future architectures should contain an AAA component that assigns each service a security profile describing the required security and trust level of the user in order to access the service (Sisalem & Kuthan, 2004)). Profiles determine the access rights for users and contain authorized users, credentials needed, certificates, subscription to groups, and so forth. Profile management can be done by the node, offering the particular service (de-centralized approach) and SM as the service advertiser when evaluating the decision to respond to queries (centralized approach). Using both centralized and de-centralized management methods, it will protect the various access networks (3G, WLAN/WMAN, PAN/PN) from unauthorized use and will keep the management procedure fast and reliable.

Context-Aware Security

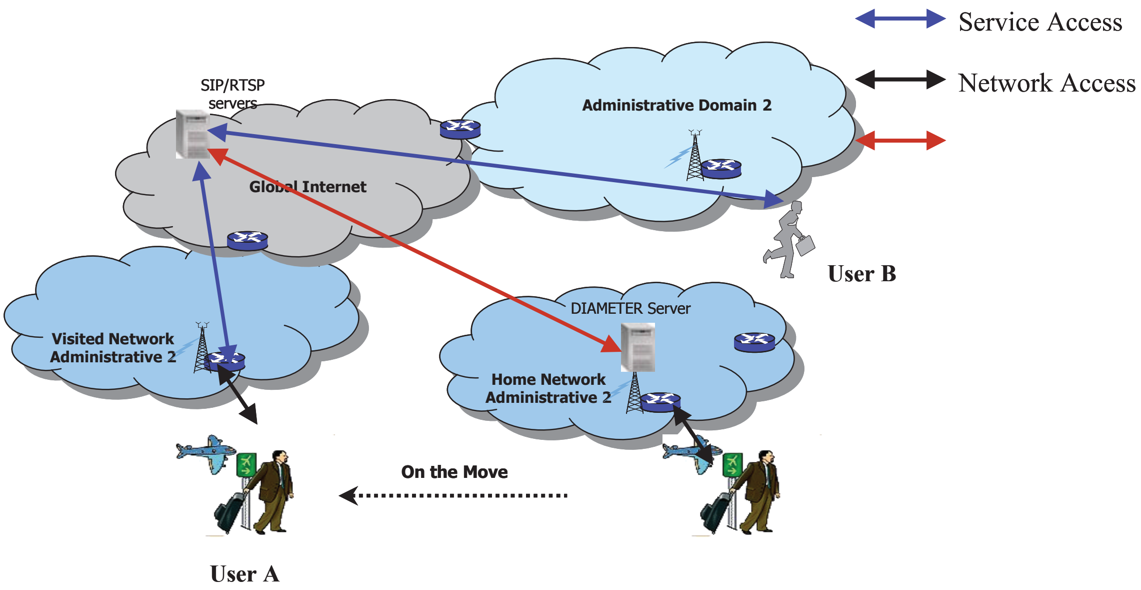
One of the basic requirements for B3G is to support seamless mobility, such that the user does not perceive any delay or interruption of service in heterogeneous networks. To provide seamless mobility, handover between networks of either heterogeneous technologies or different administrative domains must be smooth and secure. Because each network may deploy its own security mechanisms that are incompatible with others, seamless handover imposes certain restrictions for maintaining the same security level and minimum delay.

Context transfer aims to minimize the impact of certain transport/routing/security-related services on the handover performance (Loughney, Nakhjiri, Perkins, & Koodli, 2004). When a mobile node (MN) moves to a new subnet, it needs to maintain services that have already been established at the previous subnet. Such services are known as 'context transfer candidate services', and examples of these services include QoS policy, AAA profile, IPsec state, header compression, session maintenance, and so forth. Re-establishing these services at the new RAN will require a considerable amount of time for the protocol exchanges, and as a result time-sensitive real-time traffic will suffer during this time. Alternatively, context transfer candidate services state information can be transferred, for example, from the previous RAN to the new RAN so that the services can be quickly re-established. A context transfer protocol will result in a quick re-establishment of context transfer candidate services at the new domain. It would also contribute to the seamless operation of multimedia application streams and could reduce susceptibility to errors. Furthermore, re-initiation to and from the mobile node will be avoided, hence wireless bandwidth efficiency will be conserved (Georgiades, Dagiuklas, & Tafazolli, 2006).

Context transfer requirements should consider the following:

1. **Context Management and User Control:** In many activities in the research community, context information is considered to be important to support usage, operation, and management of heterogeneous wireless network (3G, WLANs, emerging 4G) and service provided by these networks.
2. **Secure Context Transfer of Security Information in Handovers Between Heterogeneous Access Technologies or Network Types:** When a handover occurs, timing constraints may forbid performing a full new access procedure, including authentication and key agreement. Instead, the security context may be transferred between points of attachment in the network which trust each other. The precise nature of the transferred security context must be specified, and the security for the discovery of points of attachment and of the transferred context need to be further studied and researched, especially during vertical handover.
3. **Secure Context Adaptation of Security Information in Handovers Between Heterogeneous Access Net-**

Figure 3. AAA and SIP interworking scenario



works: It may not be sufficient to merely transfer the security context in a handover, but the security context may need to be adapted according to the new environment. For instance, the IP address in an IPsec Security Association may change, or different cryptographic mechanisms or schemes to protect communication traffic may be used.

MULTIMEDIA PROVISIONING

The implementation of demanding services, such as real-time applications and streaming multimedia, in mobile environments using wireless connections has attracted considerable attention over the last few years. Efficient mobility management is considered to be one of the major factors towards seamless provision of multimedia applications across heterogeneous networks (Dagiuklas et al., 2005).

In order to cope with the requirements imposed by the NGN architecture, the IP multimedia subsystem (IMS) has been specified by 3GPP as a comprehensive service framework for both basic calling and enhanced multimedia services (3GPP, 2004). The IMS is the key enabler in the mobile world for providing rich multimedia services to the end users. The IMS enables complex IP-based multimedia sessions to be created with guaranteed QoS for each media component, ensuring that multimedia sessions will be able to reserve the resources they need and are authorized to use in order to perform satisfactorily. The IMS does not standardize any applications, only the service capabilities required to build various services. As a result, real-time and non-real-time multimedia services can easily be integrated over a common IP-based transport. The functions supported by the IMS include quality of service control, interworking and roaming, service control, and multiple network access. Session initiation protocol (SIP) was chosen as the session control protocol for IMS, which is based on HTTP and uses all service frameworks developed for HTTP. In addition diameter was chosen to be the AAA protocol in the IMS.

Taking into account a mixed environment with both fixed and mobile users and the currently standardized protocols SIP and diameter, several extensions for a complete and integrated security framework with a well-defined identification mechanism and a fully defined intercommunication method with AAA architecture are needed (Camarillo & Garcia-Martin, 2004). Multimedia provisioning across heterogeneous networks should consider the following aspects:

- **Application Level Security:** SIP security mechanisms can be embodied within SIP body message including encryption (enables the possibility to ensure privacy) and AAA.
- **SDP Security and Media Security Extensions:** A number of SDP extensions have been motivated by SIP-based applications, and these need to be accommodated in SDP, which now supports the use of the SRTP/SRTCP protocol. The definition of the SRTP protocol is still in draft phase and under discussion. Features associated with key management attributes need to be included (not just for SIP) and so may need to be general mechanisms to signal security capabilities.

- **SIP Identity Integrity:** The identity information asserted by the sender of a request is the 'From' header containing an URI (like 'sip: ipolitis@ee.upatras.gr') and an optional display name (like "Ilias") that identifies the originator of the request.
- **SIP Privacy:** The privacy problem is further complicated by proxy servers (also referred to in this document as "intermediaries" or "the network") that add headers of their own, such as the record-route and via headers**.**
- **End-to-End QoS:** Current approaches focus on QoS control in the access part, but the QoS control among service providers is still an open issue (Farkas et al., 2006).

CONCLUSION

In conclusion this article relates the functionalities and AAA infrastructure in order to support the dynamic establishment of trust relations between independent providers in a secure and distributed manner. The support of NGN networks involves research work on vast areas ranging from mobility, quality of service (QoS), roaming models, security, and integration with current networks. Towards the establishment of heterogeneous networks and services, solutions have been presented for supporting the provision of multimedia services over heterogeneous networks, benefiting from the availability of standardized solutions (IMS, Diameter, SIP, etc.) for supporting the operators' needs and solving issues of heterogeneity. Finally, implementing the envisaged conditions, the users' freedom to roam between different networks and use any locally available services in secure and satisfactory manner is increased.

REFERENCES

Camarillo, G., & Garcia-Martin, M. A. (2004). *The 3G IP multimedia subsystem (IMS).* New York: John Wiley & Sons.

Dagiuklas, T., & Velentzas, S. (2003). *3G and WLAN interworking scenarios: Qualitative analysis and business models.* IFIP HET-NET03, Bradford, UK.

Dagiuklas, T., Gatzounas, D., Theofilatos, D., Sisalem, D., Rupp, S., Velentzas, R., et al. (2002). Seamless multimedia services over all-IP network infrastructures: The EVOLUTE approach. *Proceedings of the IST Summit 2002* (pp. 75-78).

Dagiuklas, T., Politis, C., Grilli, S., Bigini, G., Rebani, Y., Sisalem, D., et al. (2005). Seamless multimedia sessions and real-time measurements across hybrid 3G and WLAN networks. *International Journal of Wireless and Mobile Computing,* (4th Quarter).

Farkas, K., Wellnitz, O., Dick, M., Gu X., Busse, M., Effelsberg, W., et al. (2006). Real-time service provisioning for mobile and wireless networks. *Elsevier Computer Communications, 29*(5), 540-550.

Georgiades, M., Dagiuklas, T., & Tafazolli, R. (2006). Middlebox context transfer for multimedia session support in all-IP networks. *Proceedings of the ACM Conference IWCMC,* Vancouver, Canada.

Kingston, K., Morita, N., & Towle, T. (2005). NGN architecture: Generic principles, functional architecture and implementation. *IEEE Communications Magazine,* 49-56.

Loughney, J., Nakhjiri, M., Perkins, C., & Koodli, R. (2004). *Context transfer protocol.* Internet Draft, *draft-ietf-seamoby-ctp-08.txt.*

Nakhjiri, M., & Nakhjiri, M. (2005). *AAA and network security for mobile access* (pp. 1-23). New York: John Wiley & Sons.

Salkintzis, A. (2004). Interworking techniques and architectures for WLAN/3G integration towards 4G mobile data networks. *IEEE Wireless Communications,* (June), 50-61.

Sisalem, D., & Kuthan, J. (2004). Inter-domain authentication and authorization mechanisms for roaming SIP users. *Proceedings of the 3rd International Workshop on Wireless Information Systems,* Porto, Portugal.

3GPP. (2004). *IP multimedia subsystem version 6.* 3G TS 22.228.

Velentzas, S., & Dagiuklas, T. (2005). *Tutorial: 4G/wireless LAN interworking.* IFIP HET-NET 2005, Ilkley, UK.

KEY TERMS

Authentication Authorization Accounting (AAA): Provides the framework for the construction of a network architecture that protects the network operator and its customers from attacks and inappropriate resource management and loss of revenue.

Diameter: An AAA protocol for applications such as network access or IP mobility. It is a base protocol that can be extended in order to provide AAA services to new access technologies; it is intended to work in both local and roaming AAA situations**.**

IP Multimedia Subsystem (IMS): Provides a framework for the deployment of both basic calling and enhanced multimedia services over IP core.

Liberty Alliance: Broad-based industry standards consortium developing suites of specifications defining federated identity management and Web services communication protocols that are suitable for both intra-enterprise and inter-enterprise deployments.

Quality of Service (QoS): The probability of the telecommunication network meeting a given traffic contract, or in many cases used informally to refer to the probability of a packet succeeding in passing between two points in the network.

MANET: Mobile ad-hoc network.

Next Generation Networking (NGN): A broad term for a certain kind of emerging computer network architectures and technologies which generally describes networks that natively encompass data and voice (PSTN) communications, as well as (optionally) additional media such as video.

Remote Authentication Dial-In User Service (RADIUS): An AAA protocol for applications such as network access or IP mobility.

Session Initiation Protocol (SIP): A protocol developed by the IETF MMUSIC Working Group and proposed standard for initiating, modifying, and terminating an interactive user session that involves multimedia elements such as video, voice, instant messaging, online games, and virtual reality. It is one of the leading signaling protocols for voice over IP.

Service Manager (SM): A server that discovers and manages services in its corresponding tier and interfaces with its upper-tier SM.

Triple Play: A term for the provisioning of the three services—high-speed Internet, television (video-on-demand or regular broadcasts), and telephone service—over a single broadband wired or wireless connection.

A QoS Routing Framework on Bluetooth Networking

Chao Liu
Waseda University, Japan

Bo Huang
Waseda University, Japan

Takaaki Baba
Waseda University, Japan

INTRODUCTION

Bluetooth (2001) is one of the low-bandwidth, energy-efficient wireless technologies designed for mobile devices. As the technology spreads widely in various applications, more and more services and functions are brought to the front, so different types of devices may be equipped with the Bluetooth module and appear in the same area. However, when nodes for different services come together, the need for forming a network comes out.

Actually in Bluetooth technology, there is a kind of basic form of network structure, which is called *piconet.* In a piconet of several nodes, there needs to be a master node and up to seven slave nodes, and all nodes form a star topology centered by the master node. Thus, the piconet is limited by the node number of eight and the communication range of area centered with the master node. To extend the piconet, the scatternet is proposed. In the scatternet, some slave nodes are proposed to serve more than one master. So it can act as a bridge between piconets.

However, the piconet and the scatternet is just the link layer structure. To transmit data between different piconets, in the network layer, a routing protocol is necessary. In addition, the *quality of service* (QoS) is another issue to be solved (Wang & Crowcroft, 1996), because those nodes that are equipped with a Bluetooth module could vary quite differently. Some may be multimedia data, some may be emergency data, and some may be best effort data. When all these services share the same network, the priority shall be different. For this reason, we try to propose some QoS mechanisms in the Bluetooth network.

Bluetooth has a quite unique protocol stack, which has some core layers and the other layers are left to be flexibly customized. To multiplex all above possible services, the QoS routing is proposed to be inserted directly above the core layers of Bluetooth, as shown in Figure 1.

Why shall we propose an improvement to the protocol architecture? It is one of the key points of this article. One of the reasons is as addressed above, to multiplex and not to be bypassed by upper layers. Another reason is that it can utilize the *energy efficiency* function better, which is provided by the under core layers of Bluetooth.

BLUETOOTH FEATURES AND CHALLENGES

Bluetooth is designed to be the industry standard for low-power mobile devices. Highly integrated chipsets are being developed that provide RF circuitry and protocol processing on a single chip. Current Bluetooth only supports simple services, such as human interfacing, audio/video transmitting, and home network controlling. Most of the cases are one-step communication or the last step to the wired network. Therefore, the Bluetooth network is just as simple as a star topology consisting of a master and several slaves or less. The network range is just the radio transmission range, about 30 feet. In each piconet, the master is limited to have no more than seven active slaves.

To extend the network a little bit further, a multi-hop relay is necessary. Although in Bluetooth, the *scatternet* is designed specially to build the connections between multiple piconets, it is simply a concept of link layer and physics layer, which means it offers the ability to transmit data between piconets but it does not offer the routing function. We need routing algorithms. The general routing algorithms in the wired network or the routing algorithms proposed for the mobile ad hoc networks are not efficient for the Bluetooth scatternet. The power efficiency consideration is necessary.

Bluetooth technology is designed to support 2Mbps at a range of 30 feet. The 2Mbps bandwidth is shared by all the slaves in the piconet. If several slaves try to communicate with the master, the master may share equal opportunity with each slave and the bandwidth is split. If several piconets are overlapped with each other, they will share the same radio spectrum resources as well. Therefore, in each piconet, the

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

available bandwidth also becomes less. In either case above, each link connection of the multi-hop path is trying to compete for the bandwidth with each other. Thus, if any streaming service runs over the Bluetooth network, the available end-to-end bandwidth will be quite critical. From this viewpoint, a quality of service routing is necessary.

Comparing to others, such as IEEE Standard 802.11 (1997), HyperLan (Z002), HomeRF (Chinitz, 2001), ZigBee (ZigBee Alliance, 2002), and so forth, the advantage of Bluetooth technology is that it offers the power efficiency at a low cost and relatively enough bandwidth. The Bluetooth offers the function of power efficiency by several modes of operation, which is hold mode, sniff mode, and park mode, besides the connection mode. The four modes of operation consume different amounts of energy in different activity states. The sniff mode and the hold mode are not generic device modes. Here we simply consider the connection mode and the parked mode. The connection mode is the mode that the device is working actively, while the parked mode is the mode where the device almost does not work (actually still receiving some messages from the master by a specific channel) and consumes less energy. Normally mobile ad hoc network routing protocols require that all the nodes be in the connection mode. Therefore, the advantage of Bluetooth cannot be explored. We need a routing protocol that can manage the mode of operation to save the energy by sleeping the unused nodes.

NETWORK MODELING

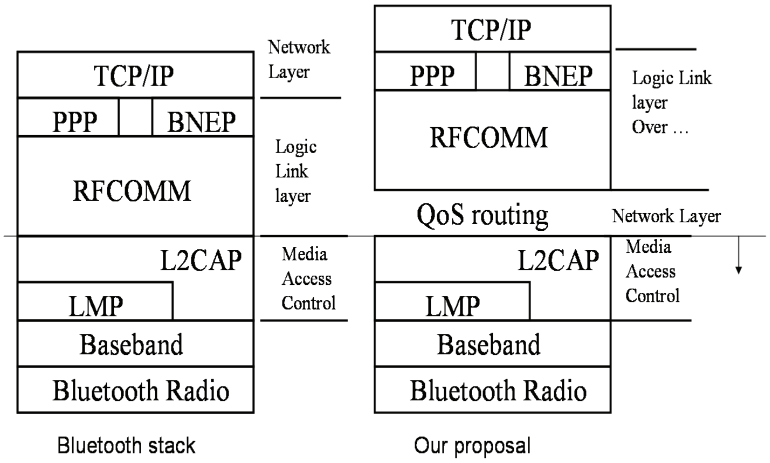
In this section, we compare different applications and different choices of the network model. As a result, it is found that routing on the L2CAP layer is the best choice.

A typical example of the Bluetooth protocol stack is shown in Figure 1. The Bluetooth radio and the baseband consist of the physical layer of the Bluetooth which is implemented by hardware chipsets. The LMP and the L2CAP layer represent the Media Access Control layer of the Bluetooth. The LMP deals with the power management and offers the ACL (Asynchronous Connectionless Link) channel for the communication. The L2CAP layer takes responsibility to multiplex logic transports into the physical channels. By these means, the upper layers can share the bandwidth. Up to the L2CAP layer, the protocols are usually integrated in the hardware. Those protocols built above the L2CAP layer are usually left to be flexibly configured according to the need. The RFCOMM (RFCOMM with TS 07.10, 1999) is protocol emulation to the serial ports. The Bluetooth Network Encapsulation Protocol (BNEP) (Bluetooth Special Interest Group, 2001) is a protocol specifically designed for the Bluetooth to emulate the Ethernet. The PPP (Simpson, 1994) is the simple and useful point-to-point protocol. According to different applications, the choice of these logic link layer protocols varies greatly.

The network model of supporting IP services over Bluetooth has been a warmly discussed issue. The possible network model can be [IP over ACL], [IP over L2CAP], [IP over [BNEP over L2CAP]], [IP over [PPP over RFCOMM]], and [IP over [PPP over L2CAP]].

The L2CAP layer has the segmentation and reassembly (SAR) function and maximum transmission unit (MTU) negotiation, which is done by hardware. Because the SAR function is an unavoidable part in the whole process, hardware implementation of course is better than additional software. Thus we prefer the [IP over L2CAP] to the [IP over ACL] network model. The other three models introduce much overhead and a lot of unnecessary things, although it can improve the compatibility to support many widely used protocols.

Figure 1. Bluetooth protocol stack



The network model of multimedia streaming over Bluetooth is also researched (Wang, 2003). Chia and Beg (2002) proposed and compared HCI, L2CAP, and IP as alternative intermediate protocols for video streaming over Bluetooth. A qualitative comparison of the three intermediate layers is made based on the size of the overheads, the efficiency of segmentation and reassembly processes, and hardware compatibility. Implementation issues of streaming video via different layers over Bluetooth are also discussed in Bilan (2003) and Scheiter et al. (2003). It suggests that video streaming via IP and L2CAP can be achieved using three Bluetooth specifications: the Local Area Network Access Point Profile – LAP, Bluetooth Network Encapsulation Specification – BNEP, and Audio/Video Distribution Transport Protocol – AVDTP (Bluetooth Special Interest Group, 2001, 2002a, 2002b).

Streaming over IP

Streaming over IP can be achieved by using LAP or BNEP. Since the stream is packetized by the IP, it has no problem streaming through the Internet. However, comparing to the streaming over L2CAP, LAP, and BNEP adds an additional encapsulation layer. Thus there are at least three encapsulation layers for an IP Bluetooth solution: L2CAP, BNEP/LAP, and IP.

Streaming over L2CAP

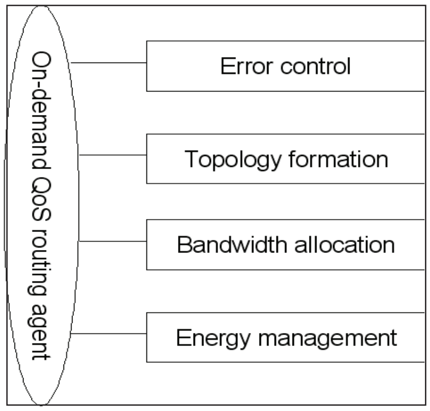
Bluetooth over L2CAP is defined by three specifications: Audio/Video Distribution Transport Protocol, Audio Video Control Transport Protocol (AVCTP), and Generic Audio/Video Distribution Profile (GAVDP). AVDTP suggests that L2CAP channels are best suited for the support of A/V stream data distribution links, because L2CAP can be flexibly configured to enable bandwidth to be shared between multiple A/V content streams.

In summary, we can see that the L2CAP is the simplest and suitable layer to make any routing function. And it shall achieve the least overhead. Although in the IP layer, the routing protocol can be implemented, it is not easy to explore the merits of the lower layer from the IP layer. The routing in the L2CAP could be the optimum choice.

QoS ROUTING

Our routing protocol lies directly above the core layers, which include the L2CAP and the LMP layer, of the Bluetooth technologies. Thus it is easy to control the Bluetooth devices. To maintain a good performance of the whole network, many aspects must be considered, such as error control, topology formation, bandwidth allocation, and energy management, as shown in Figure 2.

Figure 2. The framework of QoS routing



Bluetooth ACL link provides three built-in baseband error correction techniques:1/3 rate FEC, 2/3 rate FEC, and ARQ. The 1/3 rate FEC scheme is adopted in the header of L2CAP packets. The other two are applied to the payload. In the AVDTP, the upper-layer FEC scheme is proposed to protect video packets. The upper-layer recovery benefits from providing different protection according to the video packet types or contents. We just leave the interface to the application to select the right way of error protection.

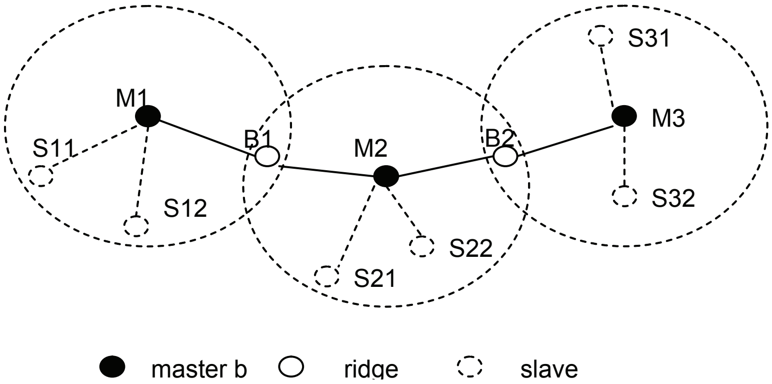
In order to sleep unnecessary nodes, our protocol will form a backbone network to maintain the routing. Other nodes may be put into the parked mode to save energy. Thus the topology formation is triggered by the services.

The bandwidth allocation is one of the key points in our routing protocol. Since all the nodes within a piconet are sharing the same bandwidth, and the master knows the current state of the piconet. It is necessary to send a request to get permission from the master when too many streams of traffic are running in the network. This is described in detail in the following sections.

Backbone Structure Formation

Forming the topology structure is the first step in making a Bluetooth network. Since energy efficiency is one of our goals, in our protocol the key roles of the network are kept awake to maintain the knowledge of the network structure. The key roles are the masters of the piconets and those bridge nodes of the scatternet. Some slave nodes are put into the parked mode when they are not sending data. As Figure 3 shows, the backbone is formed by M1, B12, M2, B23, and M3. Other nodes are parked. Because the master knows well the state of its piconet, it is enough for the master to take charge of routing information. To reduce the interference between the piconets, when forming the structure, the formed piconets need to be overlapped as little as possible.

Figure 3. The structure of Bluetooth network



Routing Identification

In the network layer, each node must have a unique identification to differentiate each other, for instance, the IP address. Here we simply use the Bluetooth device address. However, the Bluetooth device address is not put into any packet header as routing instruction, as the IP does. It is only used while routing at the beginning. After the routing path information is obtained, the tunnel is built from the source to the destination. Thus, there is no need to check the routing table for each packet. Therefore it is desired to avoid any unnecessary overhead, and we use L2CAP channels as tunnels instead of putting an ID on each packet header. As shown in Figure 4, each flow uses a tunnel. There is a default L2CAP channel, which is used as the channel of routing messages. Although there are only a limited number of L2CAP channels offered, the assumption is that we are not going to build any large-scale network or support any complicated traffic scenarios. Thus, the number of L2CAP channels is enough.

Routing Mechanisms

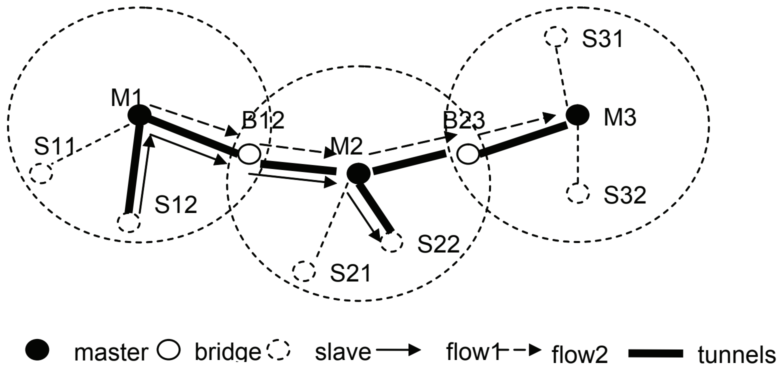
After the structure is formed, the structure information is described as in Table 1, which is the descriptive routing table of the structure in Figure 3. Each row in Table 1 is mapping to a piconet in Figure 3. For instance, the master of piconet 1 (M1) knows its member S11, S12 in parked state and this piconet has a bridge node to piconet 2 (P2). The master also knows how much bandwidth is being used, such as W1 is used in P1, and W1 is used for the bandwidth request process.

The routing procedures can be described as follows. First, as shown in Figure 4, node S12 wants to start a flow, so it wakes itself up and sends a route request message to its master that includes the information of required bandwidth from the application layer. The master checks the value of W1 to see whether there is still enough bandwidth left. If ok, then it forwards the request to the destination according to the routing table. Otherwise, it sends back a route deny message. When the bridge node B12 receives the route

Table 1. The structure information

Piconet	Used Bandwidth	Members
P1(M1)	W1	S11(Parked);S12(Parked);B12(Active)
		B12->P2
P2(M2)	W2	S21;S22;B12;B23
		B12->P1;B23->P3
P3(M3)	W3	S31;S32;B23
		B23->P2

Figure 4. The packet flows



request, the used bandwidth is checked by W1+W2 to see if the bandwidth is enough.

If the route request successfully arrives at the destination, the destination will send a route reply back. Any key role players must update their value of used bandwidth at this time. Actually the value is calculated and bound by each L2CAP channel. If any channel is broken, a route error message is sent back and the used bandwidth is updated.

CONCLUSION

In this article, we have presented a routing protocol for Bluetooth networks. Comparing to other general routing protocols, which do not consider the Bluetooth features efficiently, our protocol improves much, in that it can save energy and provide the *bandwidth reservation* function for different services, especially for multimedia services. We suggest building the routing protocol directly over the L2CAP layer of the Bluetooth. By this means, it is able to manage the mode of operation of the nodes, and energy can be saved. In our protocol, we also provide an approach to reserve the bandwidth in order to provide different QoS for different services.

ACKNOWLEDGMENTS

This work was supported by funds from the MEXT via Kitakyushu innovative cluster projects.

REFERENCES

Bilan, A.P.P.S. (2003). Streaming audio over Bluetooth ACL links. *Proceedings of the 2003 International Conference on Information Technology: Computers and Communications (ITCC'03)*.

Bluetooth. (2001). *Specification of the Bluetooth system.* Specification Volume 1.

Bluetooth Special Interest Group. (2001). *Bluetooth Network Encapsulation Protocol (BNEP) specification.* Version 0.95 Draft, Bluetooth PAN Working Group.

Bluetooth Special Interest Group. (2002a). *Audio/Video Distribution Transport Protocol specification.* Version 1.00a Draft, Bluetooth Audio Video Working Group.

Bluetooth Special Interest Group. (2002b). *Specification of the Bluetooth system: Part K:9 LAN access profile.* Version 1.1.

Chia, C.H., & Beg, M.S. (2002, December 15-18). MPEG-4 video transmission over Bluetooth links. *Proceedings of the IEEE International Conference on Personal Wireless Communications,* New Delhi.

Chinitz, L. (2001, May 9). *HomeRF technical overview.* Retrieved from www.homerf.org/data/events/past/pubseminar_0501/tech_overview.pdf

ETSI TS 101 475 V1.1.1. (2002, April). *Broadband Radio Access Networks (BRAN).* HYPERLAN Type 2, Physical (PHY) Layer.

IEEE Standard 802.11. (1997). *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications.*

RFCOMM with TS 07.10. (1999, November). Version 1.0 B.

Scheiter, C., Steffen, R. et al. (2003). A system for QoS enabled MPEG-4 video transmission over Bluetooth for mobile applications. *Proceedings of the 2003 International Conference on Multimedia and Expo (ICME '03).*

Simpson, W. (1994, July). *The Point-to-Point Protocol (PPP). Request for Comments (RFC) 1661.*

Wang, X. (2003). *Video streaming over Bluetooth: A survey.* Retrieved from http://www.comp.nus.edu.sg/~wangxia2/

Wang, Z., & Crowcroft, J. (1996). Quality-of-service routing for supporting multimedia applications. *IEEE Journal of Selected Areas in Communications, 14*(7), 1228-1234.

ZigBee Alliance. (2002). *ZigBee Working Group Web page for RF-Lite.* Retrieved from http://www.zigbee.org/

KEY TERMS

Ad Hoc: A network connection method that is most often associated with wireless devices. The connection is established for the duration of one session and requires no base station.

Flow: A sequence of continual data; the content will be anything, such as multimedia or messages.

Link Manage Protocol (LMP): Handles link control, power-sensitive states changing, and data encryption.

Logical Link Control and Adaptation Protocol (L2CAP): Used within the Bluetooth protocol stack. It passed packets to either the Host Controller Interface (HCI) or on a hostless system, directly to the Link Manager. L2CAP provides protocol multiplexing, segmentation and re-assembly, quality of service, and group addressing. Voice packets can bypass the L2CAP.

ZigBee: A published specification set of high-level communication protocols designed to use small, low-power digital radios based on the IEEE 802.15.4 standard for wireless personal area networks (WPANs).

Radio Resource Management in Convergence Technologies

G. Sivaradje
Pondicherry Engineering College, India

I. Saravanan
Pondicherry Engineering College, India

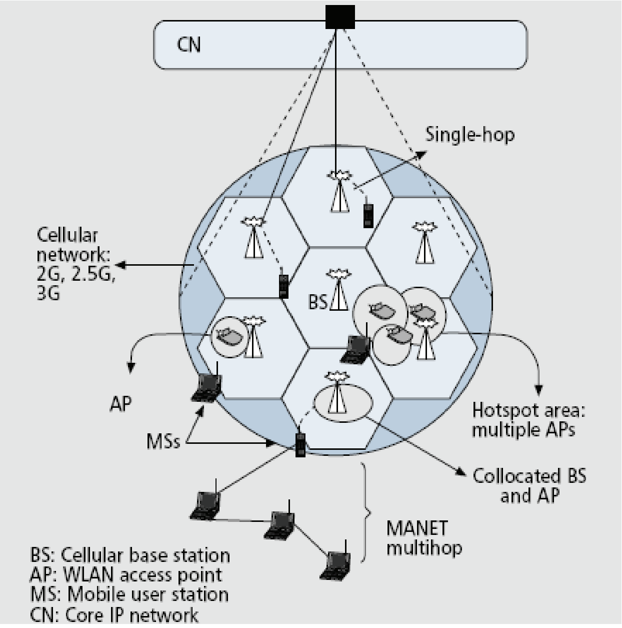
P. Dananjayan
Pondicherry Engineering College, India

INTRODUCTION

Today we find a large number of wireless networks based on different radio access technologies (RATs) and standards. Furthermore, new RATs will be developed to complement those that exist already today. Each RAT will have its strengths and weaknesses with respect to capacity, cost, achievable data rates, and support for end user mobility. As no single RAT will be able to fully support all service and user requirements, B3G networks will integrate multiple RATs in a common network. It can be anticipated that within a few years' time, a user terminal (UT) will have a choice of access technologies via which it can connect to the fixed communication infrastructure. This paves the way towards new possibilities of managing user service quality as well as radio resource utilization. The idea of multi-access radio resource management (MRRM) handling is to explore these possibilities by co-coordinating radio resource usage of different RATs such that total system capacity as well as perceived performance for individual users is increased. The integration of different technologies with different capabilities and functionalities is an extremely complex task and involves issues at all the layers of the protocol stack. The integrated heterogeneous architecture (Dave, 2005) is shown in Figure 1.

The rest of the article is organized as follows. The internetworking proposals, the challenges while approaching radio resource management, and the RRM functions are discussed. The proposals for vertical handover management are briefed, and finally multi-access RRM distribution in B3G multi-radio access networks is discussed.

Figure 1. Heterogeneous network architecture



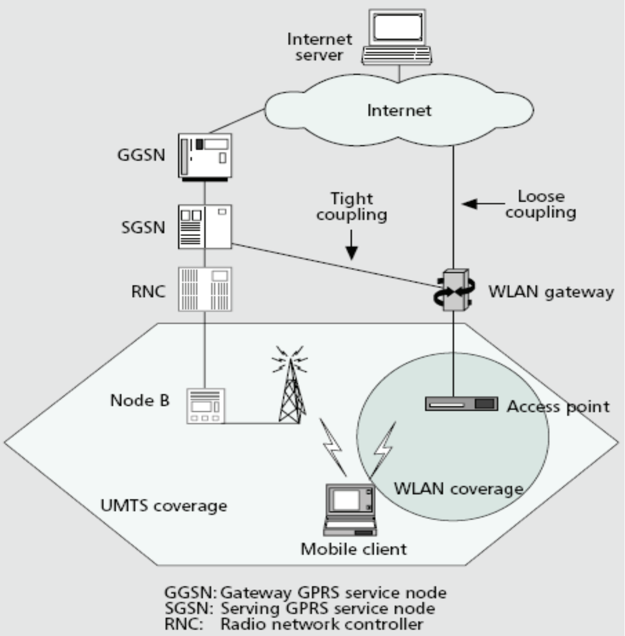
INTERNETWOKING PROPOSALS

There are three proposals for internetworking existing networks: tight coupling, loose coupling, and hybrid coupling. *Tight coupling* connects the WLAN network to the rest of the core network in the same manner as other UMTS radio access technologies. From the view of the UMTS core network, the 802.11 WLAN service area works like another GPRS serving node (SGSN) coverage area. As a result, all traffic—including data and signaling generated in the WLAN networks—is injected directly into the UMTS core network.

Loose coupling separates the data paths for the 802.11 WLAN and UMTS core networks. The WLAN gateway connects to the Internet, and all data traffic is transmitted into the core network, instead of into the UMTS core network, while signaling may optionally go through either the UMTS network or through the core Internet. *Hybrid coupling* creates a new wireless link between the base station (BS) in a cellular network and the access point (AP) in a WLAN within a same cell area using IEEE 802.16. It has advantages

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. Internetworking proposals


Internet server
Internet
GGSN
SGSN
RNC
Tight coupling
Loose coupling
WLAN gateway
Node B
Access point
UMTS coverage
WLAN coverage
Mobile client
GGSN: Gateway GPRS service node
SGSN: Serving GPRS service node
RNC: Radio network controller

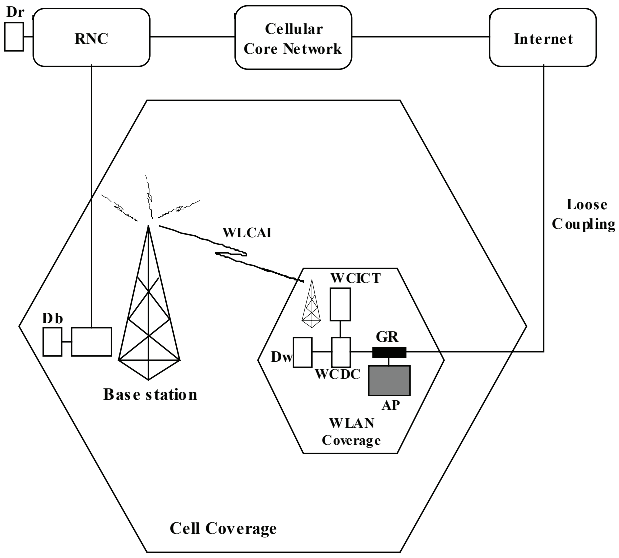
including dynamically reducing signaling cost and handoff latency due to adoption of the fast handoff techniques, and relieving the burden of core networks through dynamically distributing the traffic.

Both tight coupling and loose coupling increase the burden of core networks (Liu & Zhou, 2004), since all signaling and data transmission pass through the UMTS core network or the core Internet, and even results in bottleneck congestion or reconfiguration of network load when the interchanging traffic is too much. These two proposals are illustrated in Figure 2.

In order to overcome these shortcomings, a priority-based service interworking architecture with hybrid coupling is developed. In hybrid coupling, a new wireless link using the IEEE 802.16 standard is created between the base station (BS) in a cellular network and the 802.11 WLAN within a same cell area. Hybrid coupling has advantages including dynamically reducing signaling cost and handoff latency, relieving the burden of core networks through dynamically distributing traffic in a low-level network, and enhancing the robustness of the integrated networks through adding a new wireless link.

The priority-based service interworking architecture with hybrid coupling is presented in Figure 3. In hybrid coupling, both the cellular network and WLAN are considered as IPv6-based networks, and each element in the interworking networks has a distinct ID number corresponding to the network routing address.

Figure 3. Priority-based service hybrid coupling interworking architecture


Dr
RNC
Cellular Core Network
Internet
Loose Coupling
WLCAI
WCICT
Db
GR
Dw
WCDC
AP
Base station
WLAN Coverage
Cell Coverage

RNC	Radio Network Controller
GR	Gateway Router
AP	Access Point
WLCAI	WLAN-to-Cellular Air Interface
WCICT	WLAN-to-Cellular in Cell Transceiver
WCDC	WLAN-to-Cellular Direct Controller
Dr	Local Database in RNC
Db	Local Database in Base Station
Dw	Local Database in WLAN

RRM CHALLENGES IN FUTURE

The distribution of RRM (Magnussen, 2004) in existing RATs varies greatly, at least partly because the RATs have been optimized for different purposes. 2.5G and 3G RATs like GSM/EDGE, WCDMA, and cdma2000 have sophisticated network-centric RRM focusing on maximizing the use of available spectrum resources and also supporting mixed traffic types with different QoS requirements(voice, videoconference, Internet surfing, file download, etc.). The basic principle is that RRM decisions are made in a network node1, in most cases based on measurements collected from terminals and other network nodes. In contrast to this, IEEE 802.11 WLAN includes limited RRM support in a terminal-centric fashion, focusing on a simple (low-cost) solution providing high-peak bit rate best-effort data without any QoS. In this case the terminal is making the most of the RRM decisions, for example which access point (AP) to connect to, with some influence from parameters broadcasted by the APs. It has been identified that the simplistic RRM support in 802.11 is an obstruction to large-scale network deployment. Therefore standardization efforts are ongoing towards more efficient RRM and operation and maintenance in 802.11 networks.

Resource management schemes are strongly related to the traffic. Supporting high bit rate packet-switched traffic over the radio interface puts new requirements on resource management. Moreover, if QoS guarantees shall be provided, this will put requirements on the resource management that cannot be fulfilled by today's high data rate RATs of WLAN type. In a 4G perspective, data rates of 100 Mbit/s for wide area coverage and up to 1 Gbit/s for local coverage are envisaged. Data rates are certainly limited by propagation conditions such as multi-path and so forth, but the primary constraining factor is the terminal transmitter power, which increases linearly with the bandwidth. This means that to be able to cover large areas with high data rates, a dense infrastructure needs to be deployed. Increasing the infrastructure density may cause an increase in complexity of RRM algorithms where distributed schemes will be more and more important. Furthermore, RRM will have to operate across cells and frequencies, and may include mechanisms for dynamical reallocation and/or sharing of spectrum between operators and RATs. The side effect of using multiple RATs within the same spectrum means that the radio link will be subject to interference from other RATs, possibly in an uncoordinated manner. Reliably estimating, for instance, the carrier-to-interference ratio (C/I) as a basis for RRM decisions will be considerably more difficult. This will put requirements on future RRM functions for RATs that may co-exist within the same spectrum.

RRM FUNCTIONS

In a region with multiple independent RATs, one cannot explore the optimal use of available frequency resources and characteristics of each RAT. By coordinating the radio access networks, we can gain:

- increased trunking capacity and grade-of service (GoS) in the region;
- improved spectrum usage by selecting the best RAT based on radio conditions (e.g., path-loss);
- improved spectrum usage by selecting the most appropriate and efficient radio bearer and RAT for the service;
- minimized inter-system handover latency; and
- preservation of QoS across multiple RATs.

There are also non-performance related gains: greater flexibility for the operator with regard to infrastructure options, reduced signaling and delay due to better coordinated profile handling, and so forth.

Important MRRM functions needed to achieve these gains are:

- access discovery,
- access selection,
- admission control,
- load sharing,
- congestion control, and
- spectrum allocation.

The terminal performs *access discovery* by scanning the assigned frequencies for broadcast channels of other cells. However, to save battery power, network-assisted access discovery mechanisms may support the mobile terminal in detecting cells.

The *access selection* mechanism selects an access based on the discovered accesses, user preferences (QoS, price, etc.), and available system capacity and utilization. Two important functions within access selection are session admission control and load sharing.

Session admission control allocates resources based on QoS (and price) negotiation between the user and the network in a manner that the system load is controlled. A session admission may be followed by RAT-specific link admissions to allocate resources in each RAT.

Access can also be selected based on available capacity and utilization within different RATs—that is, *load sharing.* Intelligent load sharing mechanisms for distribution of terminals and their radio bearers between available RATs have the potential of enhancing overall system capacity.

Congestion control handles overload situations where QoS for admitted users cannot be satisfied with the normal quality agreed for a given percentage of time. The congestion control then invokes procedures that prevent some users from getting their normal QoS, preferably by first removing excess QoS (*soft congestion control*) and then disconnecting already admitted services (*hard congestion control*) based on their GoS—that is, how important different services are.

Spectrum allocation is the mechanism to (re)allocate the spectrum between multiple RATs which can be within the same or different radio access networks. MRRM functions actually needed depend on the following characteristics: support for QoS, spectral efficiency, low complexity, and so forth. The importance of these factors and the potential gain from MRRM differ between particular scenarios considered (combination of RATs, business model, etc.).

VERTICAL HANDOVER MANAGEMENT

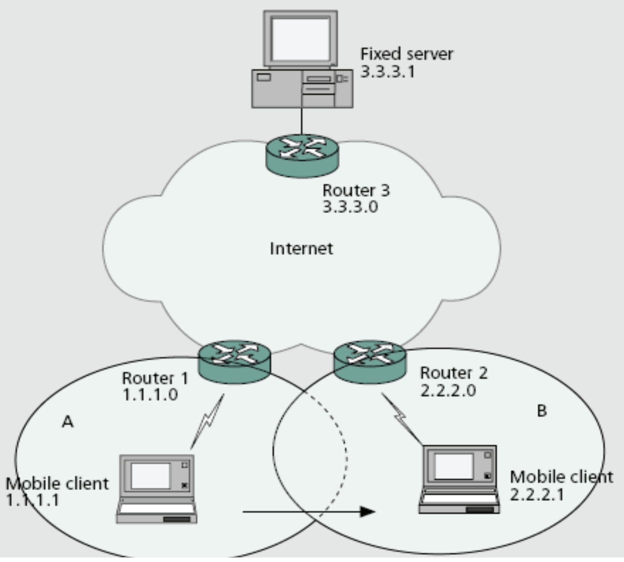
In this section the vertical handover management in convergence technologies is discussed. The movement of a user within or among different types of networks can be referred to as intersystem or vertical mobility. One of the major challenges for seamless vertical mobility is vertical handoff (Zhang et al., 2003), where handoff (or handover) is the process of maintaining a mobile user's active connections as it changes its point of attachment. Traditionally, handoff research has been based on an evaluation of the signal strength received at the mobile node, followed by a change in access point, if needed, and an updated routing path for the user connection. However, with a vision of a diverse multi-network environment, and considering the goals of transparent universal access, ubiquitous computing, and seamless mobility, traditional signal strength comparisons are not sufficient to make a handoff decision, as they do not take into account the current context or the various attachment options for the mobile user. Another issue in vertical handoff is the timely and reliable transfer of a mobile user's connection(s). While traditional link transfer techniques can achieve fast handoffs, there is now a need to consider the context of the link transfer, including security associations, QoS guarantees, and any special processing operations. Thus, the vision of 4G requires investigation of a more adaptive and intelligent network approach to vertical handoff.

In such a heterogeneous network environment, seamless mobility support is the basis of providing uninterrupted wireless services to mobile users roaming between various wireless access networks. Because of transparency to lower-layer characteristics, ease of deployment, and greater scalability, the application-layer-based Session Initiation Protocol has been considered the right candidate for handling mobility in heterogeneous wireless networks.

However, SIP (Ma, Yu, Leung, & Randhawa, 2004) entails application-layer transport and processing of messages, which may introduce considerable delay. As an application-layer protocol, SIP relies on the protocols and mechanisms in the lower layers to handle the physical network connection. As far as SIP mid-call mobility is concerned, additional procedures are needed to get the MHs attached to the wireless access network infrastructure before the SIP re-INVITE message is sent. For example, an MH attaches to the GPRS radio access network of a UMTS network using the GPRS attach and packet data protocol (PDP) context activation procedures, while it uses DHCP to attach to a WLAN. Therefore, the vertical handoff delay mainly consists of the delay of network attachment as well as that of SIP location update. In the following sections we describe the procedures of vertical handoff and analyze the associated delays. In particular, two cases of vertical handoff are of interest: handoff from a WLAN to a UMTS network, and vice versa.

A new method to facilitate seamless vertical handover between wide area cellular data networks such as UMTS and WLANs is by using the stream control transmission protocol (SCTP) (Wu, Banerjee, Basu, & Das, 2005), as shown in Figure 4.

Figure 4. SCTP support of seamless handover



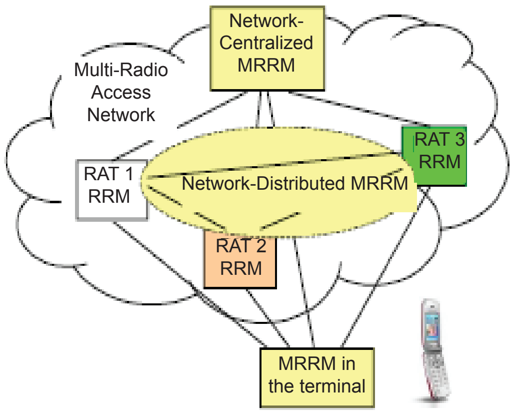
The multi-homing capability and dynamic address configuration extension of SCTP are applied in the UMTS/WLAN overlay architecture to decrease handover delay and improve throughput performance. Unlike techniques based on mobile IP or session initiation protocol, the SCTP-based vertical handover scheme does not require the addition of components such as home/foreign agents or a SIP server to existing networks. Therefore, this scheme provides a network-independent solution preferred by service providers.

MRRM DISTRIBUTION IN B3G MULTI-RADIO ACCESS NETWORKS

When distributing these MRRM functions, it is important to consider not only desired traffic characteristics but also how radio resources, corresponding measurement statistics, and control parameters are distributed throughout the network and in relation to these transport capabilities between the nodes, processing capabilities in the nodes, and so forth. From a control point of view, it is of course beneficial to place the MRRM functions where the measurement statistics and/or control originate. However, there are a number of reasons to place them elsewhere, such as:

- need for coordination of resource usage between multiple resources, such as load sharing between different cells, frequencies, RANs, and so forth;
- need for measurement statistics from multiple nodes for some decisions, such as available capacity in both uplink and downlink;

Figure 5. MRRM distribution alternatives



- CPU processing capacity needed for the RRM function;
- a transmission requirement for input parameters and control updates (periodicity, amount of data and acceptable delay);
- a specific entity wants to control the resource usage, such as the operator or the user;
- extensibility, for instance in terms of possibility for easy integration of new RATs;
- performance, for instance in terms of handover latency, spectrum efficiency, and so forth;
- scalability aspects of the different alternatives;
- possibilities for interaction with other functions, for example mobility management and general QoS control; and
- simplicity and costs, for example in terms of special software, extra nodes, and so forth.

We make a distinction here between three main categories of MRRM distribution between the terminal and the network nodes:

1. network-centralized MRRM functions,
2. network-distributed MRRM functions, and
3. MRRM functions in the terminal.

An overview illustrating these alternatives is given in Figure 5 including all possible interconnections between RRM entities.

The alternative of network-centralized MRRM means that one or more central entities are coordinating different RAT-specific RRM entities. This approach has been proposed and discussed within various forums such as 3GPP. The idea is that the non-RAT-specific MRRM functions are isolated from the RAT-specific RRM functions in one or more centralized network entities, possibly in a hierarchical structure.

In the network-distributed MRRM approach, there is no specific MRRM entity. The corresponding functionality is instead distributed between RAT-specific RRM entities through peer-to-peer relations. This principle is already today supported between WCDMA and GSM/EDGE according to the 3GPP specifications, wherein for example cell load information can be exchanged between WCDMA radio network controllers (RNCs) and GSM/EDGE base station controllers (BSCs).

The third alternative naturally means that MRRM functions and decisions are left to the terminal. This would still typically include some support from the network—that is, each network RRM entity could provide information to the terminal on which it bases its MRRM decisions.

A multi-access network may consist of a combination of these distribution alternatives, that is, some multi-access RRM functions could be centralized (e.g., overall load sharing), others could be distributed (e.g., handover of individual UTs), and some could be located in the terminal (e.g., initial RAT selection). Hierarchical coordination typically scales well. However, it may be difficult for entities high up in the hierarchy to handle detailed RRM decisions concerning individual users or cells. The idea is that the higher in the hierarchy, the less detailed and less time-critical information is handled. Furthermore, a hierarchical structure "creates" border areas where detailed coordination cannot be solved with a non-overlapping hierarchical coordination.

Distributed coordination avoids the coordination problem, but can at least theoretically run into scalability problems, as an increasing number of RRM entities in the network causes the number of interconnections between RRM entities to grow in an exponential fashion. However, in practice, geographical aspects typically limit the number of required interconnections. For instance, only each pair of RRM entities that handle overlapping or neighboring cells actually needs to have a peer-to-peer relation.

A benefit of leaving some MRRM functions to the terminal is that important measurements originate in the terminal such as link quality, and the terminal also contains the application interface, which may provide QoS requirements. In addition, it is often argued that the decision of, for example, which RAT to use should be left to the end user and not to the network operator.

This aspect is very much dependent on the involved business entities (e.g., number of operators) and underlying business model, and it may also be argued that users may not want to make such decisions, but rather just get the best possible connection for its communication needs without having to worry about the technology behind it. On the negative side for terminal control, optimal overall resource utilization is not as easily achieved as through network control. Moreover, the operator network planning becomes a very difficult task if terminal behavior is not consistent and predictable, meaning that all decision making of the terminal requires strict standardization.

CONCLUSION

This article proposes features relating RRM in convergence technologies. The convergence of all existing networks will provide access to all available services using a single-user terminal. But there are lots of challenges to be addressed in converging all networks. In spite of converging the networks, the RRM of the converged network is more challengeable. This article illustrates some of the challenges, and many more are still open issues. The complexity of radio resource management has to be addressed in the near future.

REFERENCES

Cavalcanti, D., Agrawal, D., Cordeiro, C., Xie, B., & Kumar, A. (2005). Issue in integrating cellular networks, WLANS, and MANETs: A futuristic heterogeneous wireless network. *IEEE Wireless Communications, 12*(3), 30-41.

Liu, C., & Zhou, C. (2004). HCRAS: A novel hybrid internetworking architecture between WLAN and UMTS cellular networks. In *Proceedings of IEEE 2004* (pp. 374-379).

Ma, L., Yu, F., Leung, V. C. M., & Randhawa, T. (2004). A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wireless Communications, 11*(4), 44-51.

Magnusson, P., Lundsjö, J., Sachs, J., & Wallentin, P. (2004). Radio resource management distribution in a Beyond 3G Multi-Radio access architecture. In *Proceedings of the IEEE Communications Society Globecom* (pp. 3372-3477).

Wu, W., Banerjee, N., Basu, K., & Das, S. K. (2005). SIP-based vertical handoff between WWANS and WLANS. *IEEE Wireless Communications, 12*(3), 66-72.

Zhang, Q., Guo, C., Guo, Z., & Zhu, W. (2003). Efficient mobility management for vertical handoff between WWAN and WLAN. *IEEE Communication Magazine,* 102-108.

KEY TERMS

Grade of Service (GoS): A measurement of the quality of communications service in terms of the availability of circuits when calls are to be made. Grade of service is based on the busiest hour of the day and is measured as either the percentage of calls blocked in dial access situations or average delay in manual situations.

Heterogeneous Network: A network that consists of workstations, servers, network interface cards, operating systems, and applications from many vendors, all working together as a single unit.

Radio Access Technology (RAT): Technology or system used for the cellular system (e.g., GSM, UMTS, etc.).

Radio Resource Management (RRM): Efficiently utilizing the available resources.

Wireless Local Area Network (WLAN): Wireless network that uses radio frequency technology to transmit network messages through the air for relatively short distances, like across an office building or college campus.

Wireless Metropolitan Area Network (WMAN): A regional wireless computer or communication network spanning the area covered by an average to large city.

Wireless Personal Area Network (WPAN): Personal, short distance area wireless network for interconnecting devices centered around an individual person's workspace.

Wireless Wide Area Network (WWAN): Wireless network that enables users to establish wireless connections over remote private or public networks using radio, satellite, and mobile phone technologies instead of traditional cable networking solutions like telephone systems or cable modems over large geographical areas.

RFID and Wireless Personal Area Networks for Supply Chain Management

David Wright
University of Ottawa, Canada

INTRODUCTION

Efficient supply chain management relies on knowing where products in the supply chain are located. The ability to track items from manufacturing plant to warehouse to distribution center to wholesaler to retailer is currently provided by RFID, radio frequency identification (Weinstein, 2005). Case examples of commercial applications of RFID in supply chain management are evaluated by Jones et al. (2004). A recent development, low power wireless personal area networking, WPAN, can offer advantages over RFID in certain circumstances. It is the purpose of this article to evaluate RFID and wireless personal area networks with respect to each other and to identify the features that give one an advantage over the other. We first describe the two technologies.

RADIO FREQUENCY IDENTIFICATION (RFID)

RFID tags are of two types: passive and active. A passive RFID tag is a chip incorporating memory and a microwave transmitter that is embedded in a product or in the product's packaging. The memory contains the identification number of the tag and may also contain physical specifications of the product using PML, Physical Markup Language (York, 2003). In order to read the tag an RFID reader sends out a burst of microwave energy, which is picked up by the tag and is sufficient to allow the tag to transmit the contents of its memory, which is received by the reader. Since the tag receives power from the reader, it does not need to have its own battery, and is called a passive RFID tag for that reason. Passive tags cost about US$0.20 in large volumes, and are used much more widely than active tags.

Active RFID tags incorporate a battery, cost more than passive tags and can be used to track more expensive products. The price of tags is continuously dropping and increasing usage of active tags can be expected over time.

Some tags are read-only in which case the ID is burnt into the tag at time of manufacture. Others are read/write in which case the memory contains not only the fixed identification number of the tag, but may also contain other information such as a physical description of the product (color, size, etc., in PML format), which is added when the product is manufactured.

Standardization of the identification number, so that it can be read by the many different readers used by organizations in different parts of the supply chain, started at the Auto-ID Center at MIT, and is now being pursued by EPCglobal Inc, an industry consortium that aims to standardize the format of the EPC, electronic product code for use in RFID tags. The current proposal is illustrated in Figure 1 and consists of three parts:

- A 28-bit EPC manager allowing 268 million manufacturers,
- A 24-bit object class allowing 16.8 million products for each manufacturer
- A 36-bit serial number allowing 68.7 billion copies of each product

The specification of the air interface is given by the International Standards Organization (2004). Taken together, the EPC and the air interface are the main standards for RFID.

Automated input of RFID information into a supply chain management system requires RFID readers to be located on shelves in warehouses, distributions centers, and possibly also in retail stores and in delivery trucks. Readers have a range of about one meter so that multiple readers are required. Readers can input information to the supply chain management database via wired connections, for example, using Ethernet, or using a wireless technology such as WiFi or WiMAX (Wright, 2007a, 2007b). The total cost of the system consists of the cost of the tags on each item flowing through the supply chain plus the cost of the readers. Although passive tags cost only US$0.20, readers cost approximately US$250.00.

WIRELESS PERSONAL AREA NETWORKS (WPANs)

An alternative to RFID for supply chain management is a wireless personal area network or WPAN, consisting of devices that communicate with each other instead of with a

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. 96 bit Standard Electronic Product Code, EPC

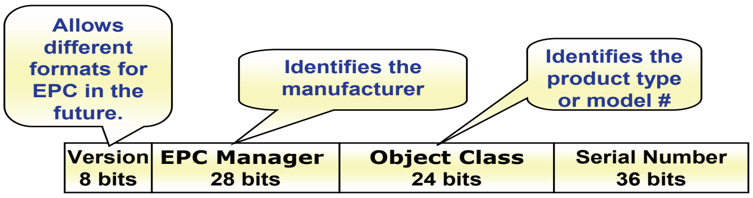
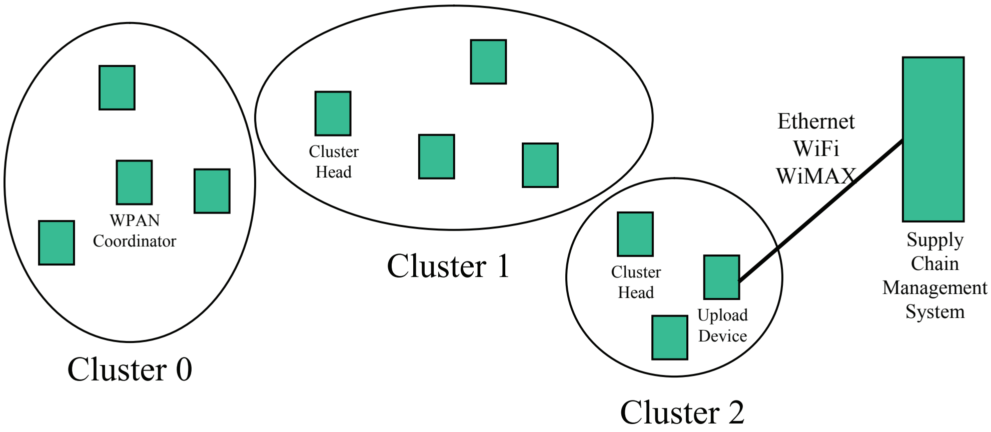

Allows different formats for EPC in the future.
Identifies the manufacturer
Identifies the product type or model #
Version 8 bits
EPC Manager 28 bits
Object Class 24 bits
Serial Number 36 bits

Figure 2. Wireless personal area network


Cluster Head
WPAN Coordinator
Cluster 0
Cluster 1
Ethernet
WiFi
WiMAX
Cluster Head
Upload Device
Cluster 2
Supply Chain Management System

reader. The word "personal" in the title does not mean that there is always a human user, instead it refers to the limited range of the wireless communications: approximately 1 meter from one device to another. WPANs are of various types and here we focus on the low power version that is standardized by IEEE (2003), and is being commercialized by the industry consortium, the Zigbee Alliance, which has developed a specification for wireless personal area network applications (Zigbee, 2006). WPANs require each device to be powered, typically with a battery, but they transmit low data rates at low power so that battery life can exceed a year. Methods for reducing power requirements are described by Liang (2003) and Rajendran et al.(2006); and the system's performance is analyzed by Chin et al. (2003). Initially applications of low power wireless personal area networks include interactive toys and industrial control, in which sensors measure temperature, humidity and position of items in a production facility and use the WPAN to communicate this information to a production control system (Egan, 2005). Supply chain applications include embedding WPAN devices with EPCs in products or their packaging and could also include sensors for measuring temperature and humidity, which are important for perishable items such as produce. WPAN devices can be designed as small as a coin, so as to be easily embedded in products and packaging (Choi, 2003).

The network architecture of a WPAN is illustrated in Figure 2, and is built up of clusters of devices. Each cluster has one device designated as a "cluster head," and one of these cluster heads is the WPAN coordinator. Although the communications range between one device and another is limited to about one meter, networks of clusters can communicate over a much longer distance. Each device uses wireless communications to communicate with its neighbors, and one device can communicate with a supply chain management system using Ethernet, WiFi or WiMAX. This "upload device" is more costly than the others, however only one is required per WPAN. Actual $ costs are not available at the time of writing (1Q06) since WPAN devices are at an early stage of commercialization. Practical methods for designing WPAN networks are described by Minami (2004).

In supply chain management, WPAN devices can be embedded in products or in packaging and can communicate with each other in warehouses and delivery vehicles. As items are added to or removed from a stack of shelves in a warehouse, the clusters reconfigure dynamically, and information about the new products is distributed to the

Table 1. Comparative evaluation of RFID and WPANs for supply chain management

	RFID	WPAN
Communications range	~ 1 meter: tag to reader.	~ 1 meter: device to device.
Networking	N/A	A network can consist of hundreds of devices.
Number of devices needed	One tag per product or package. One reader within a meter of each product or package connected to the SCM system.	One device per product or package. One upload device per WPAN, connected to the SCM system.
Cost	Very low cost per tag. Higher cost per reader.	Low cost per device. Higher cost per upload device.
Environmental monitoring	No standard integration with sensor technology.	WPAN sensors commercially available.
Reliability	< 100% reading ratio.	Error control protocol enhances reliability.
Stage of development	Established technology.	New technology at early stage of commercialization.
Power	Tags derive power from reader.	Devices need batteries or mains power.

upload device. Each stack of shelves in a warehouse needs an upload device connected to the supply chain management system to forward information it receives from all the other devices on the products stored on those shelves.

COMPARATIVE EVALUATION

A comparative evaluation of RFID and WPANs for supply chain management is summarized in Table 1.

Both RFID and WPANs have a limited communications range of about 1 meter, however the extent of a WPAN can be much greater end to end. In warehousing applications, this allows a WPAN to span an entire stack of shelves, using only a single upload device, whereas multiple RFID readers are required, each having an upload capability. Upload devices and RFID readers cost more than regular WPAN devices and RFID tags, so that using less of them gives a potential cost advantage to WPAN over RFID. Another advantage of WPAN is that additional sensor devices can be added to monitor temperature and humidity, which is important in the case of perishable products. This can be done in a standardized way with WPAN sensors; however sensors interfaced to RFID tags require a proprietary interface (Philipose et al., 2005). The final advantage of WPAN is improved reliability, due to the use of an error control protocol. Any low power wireless communication can be unreliable in the presence of metal, for example, in warehouse shelving and in the products themselves, and read ratios on RFID readers are often <100%. The downside to WPANs is the additional cost of the devices, first due to the fact that it is a relatively new technology, and second due to the requirement for each device to have a battery or mains power supply.

CONCLUSION

RFID is a mature technology that is currently seeing widespread deployment to provide information for supply chain management. WPAN is a more recent development which could provide similar and, in the case of perishable products, improved functionality. As WPAN technology is commercialized and unit prices drop, we can expect cost advantages compared to RFID, since the cost of multiple upload devices can be eliminated in the case of WPANs. In addition WPAN is a more reliable technology than RFID since it incorporates error control.

REFERENCES

Chin, F., Zhi, W., & Ko, C-C. (2003). System performance of IEEE 802.15.4 low rate wireless PAN using UWB as alternate-PHY layer. *Personal, Indoor and Mobile Radio Communications*, *1*, 487-491.

Choi, P., Park, H.C., Kim, S., Park, S., Nam, I., Kim, T.W., Park, S., Shin, S., Kim, M.S., Kang, K., Ku, Y., Choi, H., Park, S.M., Lee, K. (2003). An experimental coin-sized radio for extremely low-power WPAN (IEEE 802.15.4) application at 2.4 GHz. *IEEE Journal of Solid-State Circuits*, *38*(12), 2258-2268.

Egan, D. (2005) The emergence of ZigBee in building automation and industrial control. *Computing & Control Engineering Journal*, *16*(2), 14-19.

International Standards Organization. (2004). *ISO 18000, Radio frequency identification for item management. Parameters for air interface communications.*

IEEE. (2003). *802.15.4 low rate and ultra low power wireless personal area network*.

Jones, P., Clarke-Hill, C., Shears, P., Comfort, D., & Hillier, D. (2004). Radio frequency identification in the UK: Opportunities and challenges. *International Journal of Retail & Distribution Management*, *32*(2/3), 164.

Liang, Q. (2003). A design methodology for wireless personal area networks with power efficiency. *Wireless Communications and Networking*, *3*, 1475-1480

Minami, M., Saruwatari, S., Kashima, T., Morito, T., Morikawa, H., & Aoyama, T. (2004). Implementation-based approach for designing practical sensor network systems. In *11th Asia-Pacific Software Engineering Conference* (pp. 703-710).

Philipose, M., Smith, J. R., Jiang, B., Mamishev, A., Roy, S., & Sundara-Rajan, K. (2005). Battery-free wireless identification and sensing. *IEEE Pervasive Computing*, *4*(1), 37-45.

Rajendran, V., Obraczka, K., & Garcia-Luna-Aceves, J. J. (2006). Energy-efficient, collision-free medium access control for wireless sensor networks. *Wireless Networks*, *12*(1), 63.

Weinstein, R. (2005). RFID: A technical overview and its application to the enterprise. *IT Professional*, *7*(3), 16-21.

Wright, D. (2007a). Wireless technologies for mobile computing and commerce. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce*. Hershey, PA: Idea Group Reference.

Wright, D. (2007b). Business and technology issues in wireless networking. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce*. Hershey, PA: Idea Group Reference.

York, C. (2003). Auto ID in the supply chain of today and tomorrow. *Material Handling Management*,*58*(11), 33-35.

Zigbee. (2006). Zigbee specification. Retrieved March 2006, from http://www.zigbee.org/en/spec_download/download_request.asp.

KEY TERMS

Electronic Product Code (EPC): A code that can uniquely identify each item in the supply chain, by specifying the manufacturer, the product code and the serial number of each item.

Physical Markup Language (PML): A language for describing the physical characteristics of items, including the dimensions, weight and operating specifications of products in the supply chain.

Radio Frequency Identification (RFID): A system consisting of tags containing identification numbers, which can be transmitted to readers over a distance of about 1 meter.

WiFi: A commercial implementation of the IEEE 802.11 standard for wireless communications up to about 100 meters, in which the equipment has been certified by the WiFi Alliance, an industry consortium.

WiMAX: A commercial implementation of the IEEE 802.16 standard for wireless communications with a range of 2-5 Km, in which the equipment has been certified by the WiMAX Forum, an industry consortium.

Wireless Personal Area Network (WPAN): A network consisting of multiple devices communicating using wireless, each within about 1 meter of its neighbor, following standards developed by IEEE 802.15.

Zigbee Alliance: An alliance of companies commercializing WPAN technology by developing specifications for its efficient application in a number of areas of business, home and industry.

Scatternet Structure for Improving Routing and Communication Performance

Bo Huang
Waseda University, Japan

Chao Liu
Waseda University, Japan

Takaaki Baba
Waseda University, Japan

INTRODUCTION

As a new promising short-range wireless technology, Bluetooth (Bluetooth; Bray & Sturman, 2001; Miller & Bisdikian, 2000) is designed to enable voice and data communication among various devices. It has received great attention in recent years. Bluetooth SIG develops Bluetooth specifications. By using the unlicensed 2.4 GHz ISM band, Bluetooth devices can communicate within a local area, intended as a replacement of interconnect cable. It supports connection-oriented and connectionless links and thus is suitable for both voice and data communications. The characteristics of low cost and low energy consumption make it not only the ideal technology for wireless local area network but also the best candidate for wireless personal area network (Haartsen, 1998).

When two Bluetooth devices communicate with each other, they must set up a link at first. One device acts as master and the other acts as slave. Actually, Bluetooth specifications can support up to seven slaves for one master. The mini-network constructed by one master and several slaves is called piconet.

Limited to the communication range and slave number, usually, a single piconet is not enough for actual usage. Also, if we permit Bluetooth devices in the same short range to communicate freely, they will influence each other and lower down the whole performance. Fortunately, one Bluetooth device can act as slave in several piconets and master in one piconet simultaneously. Such Bluetooth devices, existing in several piconets, are called relay. Through relay, several piconets can connect and form a scatternet. Scatternet not only provides larger range communication, but also regulates Bluetooth devices' actions to improve the whole performance.

To construct a scatternet, the structure and routing are two major issues. The scatternet structure deals with automated procedures to select master and slaves in a piconet and relays between piconets, and the topologic structure of scatternet to improve scatternet performance. The routing algorithm deals with delivery of messages in such a scatternet (Sairam, Gunasekaran, & Redd, 2002).

Many scatternet structures and routing algorithms have been proposed in recent years. The following paragraphs will analyze three scatternet structures. After that, we present a new topologic structure, named SolidRing, with its routing algorithm. Finally, there are some discussions on SolidRing's performance.

BACKGROUND

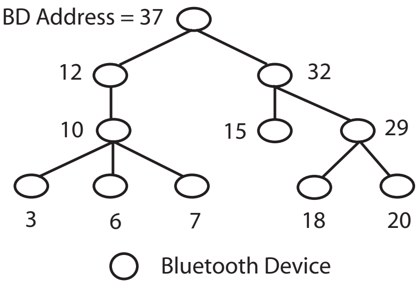
Usually, when designing scatternet, the two issues, structure and routing algorithm, can't be divided absolutely, and influence each other. With a suitable routing algorithm, scatternet performance would be remarkably improved. Of course, to different structures, the meaning of "suitable" is very different. Also, a specific routing algorithm can only be realized in some structures (Prabhu & Chockalingam, 2002; Bhagwat & Segall, 1999; Shih, Wang, & Su, 2003). In a scatternet, the two should be designed appropriate to each other.

In recent papers of scatternet, various structures and routing algorithms are proposed. After studying the papers, two basic routing algorithms and three basic structures can be concluded.

Generally, basic routing algorithms are: proactive and reactive (Royer & Toh, 1999). Reactive routing algorithms find routes only when needed, while proactive algorithms maintain valid routes all the time. So, proactive consumes more energy and memory to win time while reactive contrary. Since Bluetooth initially was designed as a low price and low energy technology, a small Bluetooth device can't accommodate the large table and power consumption needed for proactive. Reactive is not an advisable choice as well. Because Bluetooth can only communicate through comparatively narrow master-slave links, routing data flow and long waiting time can't be tolerant. For our new structure, SolidRing,

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Tree structure



we will present a new routing algorithm, a combination of the two. The routing algorithm is simple and fast, but it can only execute in SolidRing for the structure has some special characteristics (Kim, Lai, & Arora, 2003).

THREE STRUCTURES

For scatternet structure, there are three basic types: tree structure, planar structure and ring structure. Each structure has its own merits and demerits.

Tree Structure

The topologic link of tree structure is shown as Figure 1. The merit of tree structure is a very simple routing algorithm, just as described in Sun, Chang, & Lai (2002). The structure is constructed to be regular according to BD Address (Bluetooth Device Address). Root node is max BD Address device of the scatternet. Its sub-trees have different BD Address ranges, for example, the 1st sub-tree has the range of 100-200, the 2nd sub-tree has the range of 200-500, the 3rd sub-tree has the range of 500-700, and so on. BD Address ranges of sub-trees will not overlap each other. Following the rule, each node has several sub-trees whose BD Address ranges don't overlap. And, each node keeps information about its sub-trees BD Address range.

When a node tries to communicate with another node, it will compare the destination BD Address with its record of BD Address ranges of its sub-tree. If the destination BD Address is out of the range, it will pass a routing message to its father node. Otherwise, it will pass a routing message to the corresponding sub-tree. Each node that accepts the routing message, will act like this. Following the algorithm, the message will arrive at the destination. Then, the message will send back and the routing work finished.

By analyzing the routing algorithm clearly, we can find that the routing message is directly passed from start node to destination node without any unnecessary message. It's a very efficient routing algorithm and should owe to the regular formation. Routing work is sharply reduced. While the structure has an excellent routing performance, its demerits are critical.

The first problem is serious bottleneck. The structure looks like a pyramid. From bottom to top, the paths become less and less. Even on the top of the pyramid, all communications must pass through root node of the tree. Obviously, it will result in serious bottleneck. The following will give some mathematic calculations to show the bottleneck problem.

Assuming there is a b-branch, n-layer tree and all nodes in the tree communicate with each other once. From root node, the tree can be divided to number of b sub-tree, that each one is a branch with its own sub nodes. The communications related to a sub-tree include two parts: internal and external communications. Internal means the sub-tree nodes communicate with each other. External means the sub-tree nodes communicate with nodes that don't belong to the sub-tree. Then, for the sub-tree, the communications that must pass through root node are just the external communications. The total nodes number is N, and a sub-tree nodes number is B. Then, we can calculate the percent of communications passing through root node.

The total nodes number

$$N = b^n - 1 \qquad (1)$$

A sub-tree nodes number

$$B = b^{n-1} - 1 \qquad (2)$$

Sum of communications of assumption

$$P_N^2 \qquad (3)$$

Total communications for a sub-tree

$$B(N-1) \qquad (4)$$

Internal communications

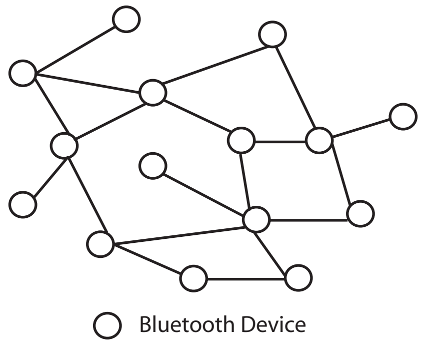
$$P_B^2 \qquad (5)$$

Then, percent of communications passing through root node is

$$b \times \frac{B(N-1) - P_B^2}{P_N^2} \times 100\% \qquad (6)$$

According to the equation (6), we can calculate that 66.7% communications have to pass root node for 3-branch, 5-layer tree and 51.6% communications for 2-branch, 5-layer tree. That means most of communications must pass through root node. In such situation, root node is very easy to be blocked and bottleneck problem is very serious.

Figure 2. Planar structure



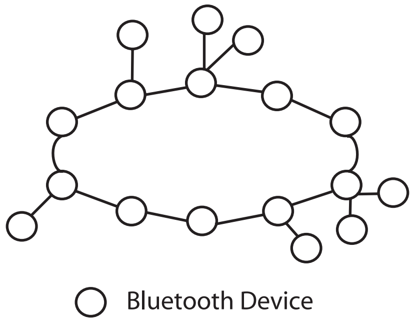
The second problem is that the communication path is so long. There are some comparison results described in Wang, Thomas, and Haas (2002). Although a tree structure could equip with a remarkable routing algorithm, it can't be the main body of a scatternet.

Planar Structure

The topologic link of planar structure is shown as Figure 2. Many papers have proposed various scatternet structures, for example, Wang, Thomas, and Haas (2002) and Wang, Stojmenovic, and Li (2004), and all of them can be concluded to be planar structure. The merits of planar structure are shorter communication paths and some immunity to the bottleneck problem. With the same device number, the layer number of a planar structure is less than those of tree and ring structures, which results in shorter communication paths. Usually, the structure has a multi-path from one node to another. This characteristic makes the structure immunity to bottleneck. In fact, these merits should owe to various rings inside the structure.

As the structure is usually constructed by random algorithms and formed an irregular structure, routing work is complicated and troublesome (Racz, Miklos, Kubinszky, & Valko, 2001). Usually, two basic routing algorithms, proactive and reactive, are often used in planar structure. Reactive routing algorithms find routes only when needed while proactive algorithms maintain valid routes all the time. So, proactive consumes more energy and memory to win time while reactive contrary. Since Bluetooth is initially designed as a low price and low energy technology, a small Bluetooth device can't accommodate the large memory and power consumption needed for proactive. Reactive is not an advisable choice as well. Because Bluetooth can only communicate through comparatively narrow master-slave links, routing data flow and long waiting time would seriously influence WPAN performance. Both algorithms are not suitable for WPAN. As the conclusion, planar structure should be the main body of a scatternet, but it must be improved.

Figure 3. Ring structure



Ring Structure

The topologic link of ring structure is shown as Figure 3. If a scatternet were constructed as a ring, the layer number would be too large. Long communication paths and bottleneck both exist. But, in some situations, a ring will reduce communication paths as well (Lin, Tseng, Chang, & Tu, 2002). For example, if the head node in a chain communicates with the tail node, the communication must pass through the whole chain. Suggesting connect the head node with tail node to transfer a chain to ring structure, the communication path length will sharply reduced to 1. The matter happens not only for head and tail nodes, but also for the whole ring.

Ring structure can also provide various paths from one node to another by adding to other structures (Wang, Thomas, & Haas, 2002). Without rings, planar structure is just some type of tree structure. Although ring structure couldn't be the main body of a scatternet, it would play an important role in scatternet.

SOLIDRING STRUCTURE

To construct the new structure, some assumptions are made regarding the environment considered herein.

1. All Bluetooth devices are within a range that enables a connection to be established. That means any two devices can receive signals sent out by each other.
2. In the initial stage, no link existed between any two Bluetooth devices.

When designing the new structure, we also take the following notices into account:

1. Avoid forming further piconets inside a piconet.
2. Avoid setting up more than one connection between two piconets.
3. Inside a piconet, the master should have suitable number of slaves.

Figure 4. SolidRing structure

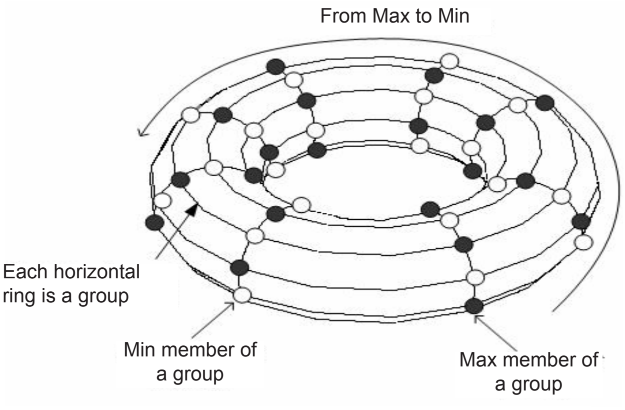
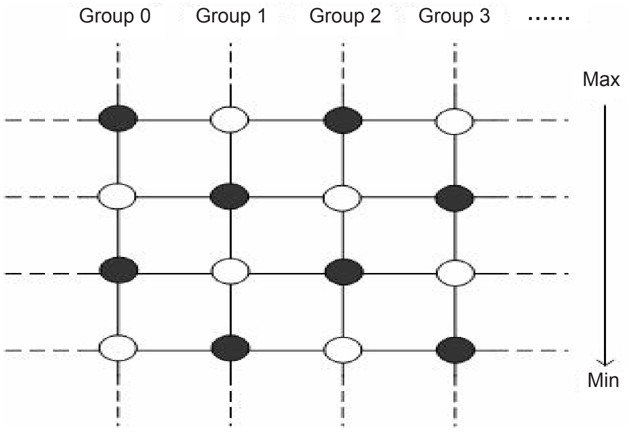


Figure 5. SolidRing section



4. Reduce the number of piconets that share a common node.

In fact, the new structure originates from tree structure. As the tree structure has excellent performance for routing, we thought of improving it to avoid bottleneck. From the description about tree structure, a 2-branch tree is better than 3-branch tree on the bottleneck problem. It can be concluded that the less the branch number in a tree, the less the bottleneck problem. Then, a 1-branch tree seems better than 2-branch tree. 1-branch tree is just a chain. So, a regularly formed chain emerges.

But, if a tree is reformed to be a chain, the layer number will increase and communication path length emerges to be a big problem. In the background section, we have concluded that planar structure has the genius to be the main body of a scatternet structure for its immunity to bottleneck problem and shorter path length. So, we thought of using numbers of such chains to weave a planar structure. To do that, the Bluetooth devices must be assigned to several groups.

For convenience of routing and communication, Bluetooth device assignment should be according to BD Address. Each device divides its own BD Address by group number N and decides which group it should join. The group number N is decided initially. BD Addresses are divided by N and get residual 0, 1, …, N-1. The residual decides which Bluetooth device belong to which group.

After grouping, each group will form a chain following the sequence from larger BD Address to smaller BD Address. The operation reduces members in a chain.

But, from max member to min member in a group, almost the whole chain must be passed through. By using ring structure, the path length in same chain reduces partially. If max BD address member connected to min BD address member, the path length between the two will be only 1. The whole length between members in same group will reduce to about half of original. Then, group formation is transformed from chain to ring.

Now, there are N groups, which means N Rings. To form as planar structure, each group ring should connect with its two neighbors. Group 1 connects with Group 0 and Group 2, Group 2 connects with Group 1 and Group 3, …, Group N-2 connects with Group N-3 and Group N-1, and Group 0 connects with Group N-1 and Group 1. The connecting rule is the max members of neighbor groups connect with each other, and along each ring, from max to min, members of neighbor groups will connect one by one.

After all groups connect with each other following the rule, the SolidRing structure emerges, just as shown in Figure 4.

If we split the SolidRing structure, its section will be as shown in Figure 5: just a Planar structure.

ROUTING ALGORITHM

In the SolidRing scatternet, each node records its four (or less) neighbors' BD Address. In the structure, the routing algorithm is very simple and fast. When a node communicates with another one, it divides the destination BD Address by group number and gets the residual. Then, it is clear which group the destination belongs to. The routing message can be directly passed to the destination group. When a routing message arrives to a destination group, no matter which node in the group receives the message, it will pass the message along the group ring. Of course, passing direction is decided by comparison of destination's BD Address and its own BD Address. If the destination exists, the destination can be found sooner or later along the group ring and routing is successful. If not, after circling along the ring, the message will return to the group node that initially received the message. That means the destination doesn't exist and routing fails. The fail message will be returned. No matter success or fail, the routing message is sent just following a single path. The routing data is sharply reduced and the bandwidth is efficiently used.

It is just a simple example for the routing algorithm. Actually, if we do some modifications on the routing algorithm, the routing procedure will select an idle route to the destination and avoid communication concentrate in some block of the scatternet, which would prevent the scatternet from bottleneck problem. The structure even has the potential to realize multi-path communication.

To realize these functions, we must do some modification for each node. That is, each node must record its own bandwidth occupation ratio during its work. When a new route is setup and asks passing through the node, the node should check its own bandwidth occupation ratio and judge whether it can work for the route. The detailed description is given as following:

- **Step 1:** Each node should record its own bandwidth occupation ratio, which will help nodes to judge whether itself is crowded or not.
- **Step 2:** When start node try to communicate with destination node, it must calculate how much bandwidth it will occupy. Then, it sends the request of communication through the whole path that has been found according to the first routing algorithms.
- **Step 3:** Each node that occurs in the original path will compare its own idle bandwidth with the request bandwidth. The compare result will be recorded inside the request message and then be returned to the start node.
- **Step 4:** After receiving the feedback message, the start node will judge along the original path which node can work for this communication and which node can not. Then, it will send another message to ask nodes along the original path to modify the path.
- **Step 5:** Each node that can not work for this communication will try to find another node or another short path to replace itself. It will send request message around to its neighborhood to ask them to take over the job. If one neighbor is not so crowded, it will take place the node and form a new path.
- **Step 6:** After all crowded nodes find their replace node or path, a new path is generated. Each node along the new path should update its own record on bandwidth occupation. Then, the communication will be carried out along the new path.

Based on this routing algorithm, communication can always choose an idler path. If we do more modification for the routing algorithm, it is also easy to realize multiple-path communication.

CONCLUSION

In the article, we analyze Bluetooth scatternet and conclude which issues are important for constructing scatternet. Based on mathematics, we describe merits and demerits for existing scatternet structures and their routing algorithms. Then, we present a new scatternet structure, named SolidRing, its construction procedure and routing algorithm. We believe SolidRing structure can work better than existing scatternet structures.

ACKNOWLEDGMENTS

This work was supported by fund from the MEXT via kitakyushu innovative cluster project.

REFERENCES

Bhagwat, P., & Segall, A. (1999). A routing vector method (RVM) for routing in Bluetooth scatternets. In *Proceedings of IEEE International Workshop on Mobile Multimedia Communications (MoMuC 1999)* (pp. 375-379).

Bluetooth Special Interest Group. (n.d.). *Bluetooth Core Specification, version 1.2.*Retrieved from http://www.bluetooth.com/

Bray, J., & Sturman, C. F. (2001). *Bluetooth: Connect without cables*. Prentice Hall.

Haartsen, J. (1998). BLUETOOTH: The universal radio interface for ad hoc wireless connectivity. *Ericsson Review*, *3*, 110-117.

Kim, Y. M., Lai, T. H., & Arora, A. (2003). A QOS-aware scheduling algorithm for Bluetooth scatternets. In *Proceedings of the 2003 International Conference on Parallel Processing*. New York: IEEE.

Lin, T.-Y., Tseng, Y.-C., Chang, K.-M., & Tu, C.-L. (2002). Formation, routing, and maintenance protocols for the BlueRing scatternet of Bluetooths. In *Proceedings of the 36th Hawaii International Conference on System Sciences*. New York: IEEE.

Miller, B. A., & Bisdikian, C. (2000). *Bluetooth revealed: The insider's guide to an open Specification for global wireless communications*. Prentice Hall.

Prabhu, B. J., & Chockalingam, A. (2002). A routing protocol and energy efficient techniques in Bluetooth scatternets. In *Proceedings 2002 IEEE International Conference on Communications (ICC 2002)* (Vol.5, pp. 3336-3340).

Racz, A., Miklos, G., Kubinszky, F., & Valko, A. (2001). A pseudo random coordinated scheduling algorithm for Bluetooth scatternets. In *Proceedings of ACM MobiHoc 2001* (pp. 193-203).

Royer, E.M., & Toh, C.K. (1999). A review of current routing protocols for ad hoc mobile wireless networks. *IEEE Personal Communications*, *6*(2), 46-55.

Sairam, K., Gunasekaran, N., & Redd, S. R. (2002). Bluetooth in wireless communication. *IEEE Communications*, *40*(6).

Shih, K.-P., Wang, S. S., & Su, J.-H. (2003). A Bluetooth group-scatternet formation algorithm for efficient routing. In *Proceedings of the 2003 International Conference on Parallel Processing Workshops*. New York: IEEE.

Sun, M.-T., Chang, C.-K., & Lai, T.-H. (2002). A self-routing topology for Bluetooth scatternets. In *Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks*. New York: IEEE.

Wang, Z., Thomas, R. J., & Haas, Z. (2002). Bluenet: A new scatternet formation scheme. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* New York: IEEE.

Wang, Y., Stojmenovic, I., & Li, X.-Y. (2004). Bluetooth scatternet formation for single-hop ad hoc networks based on virtual positions. *IEEE*, pp. 170-175.

KEY TERMS

Bluetooth: A short-range wireless communication technology. The communication range is about 10 meters. The characteristic is low power-consumption. Now, it’s widely used in our life.

Bluetooth Device Address: For each Bluetooth device, it has a unique address, which is named Bluetooth device address.

Piconet: The mini-network formation for Bluetooth devices. It can include up to 8 Bluetooth devices. One acts as master and the others act as slaves.

Scatternet: The extended network formation for Bluetooth devices. It can provide extended communication range and include more than 8 Bluetooth devices.

Tree Structure: A common network topologic structure. For Bluetooth scatternet, it’s a regularly formed structure.

Planar Structure: a common network topologic structure. For Bluetooth scatternet, it’s an irregularly formed structure.

Ring Structure: A common network topologic structure. For Bluetooth scatternet, it’s an irregularly formed structure.

Solidring Structure: Our newly developed scatternet structure. It is formed according to Bluetooth device address regularly. We believe solidring structure can work better than existing scatternet structures.

Secure Agent Data Protection for E-Commerce Applications

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

One hindrance to the widespread adoption of mobile agent technology (Johansen et al., 2002) is the lack of security. SAFER, or Secure Agent Fabrication, Evolution, and Roaming, is a mobile agent framework that is specially designed for the purpose of electronic commerce (Zhu, Guan, Yang, & Ko, 2000; Guan & Yang, 1999, 2003; Yang & Guan, 2000). By building strong and efficient security mechanisms, SAFER aims to provide a trustworthy framework for mobile agents. While such an agent transport protocol provides for the secure roaming of agents, there are other areas related to security to be addressed.

Agent integrity is one such area crucial to the success of agent technology. The integrity protection for agent code is relatively straightforward. A more complex code integrity scheme to handle code-on-demand is also proposed in Wang et al. (2002). Agent data, however, is dynamic in nature and will change as the agent roams from host to host. Despite the various attempts in the literature (Chionh, Guan, & Yang, 2001), there is no satisfactory solution to the problem so far. Some of the common weaknesses of the current schemes are vulnerabilities to revisit attack and illegal modification (deletion/insertion) of agent data.

DESCRIPTION OF SADIS

SADIS has been designed based on the following assumptions:

1. Entities including agents, agent butlers, and hosts should have globally unique identification number (IDs).
2. Each agent butler and host should have a digital certificate that is issued by a trusted CA. These entities will be able to use the private key of its certificate to perform digital signatures and encryption.
3. While the host may be malicious, the execution environment of mobile agents should be secure and the execution integrity of the agent can be maintained.
4. Entities involved are respecting and cooperating with the SADIS protocol.

Key Seed Negotiation Protocol

The proposed key seed negotiation protocol defines the process for key seed negotiation, as well as session key and data encryption key derivation. When an agent first leaves the butler, the butler generates a random initial key seed, encrypts it with the destination host's public key, and deposits it into the agent before sending the agent to the destination host. It should be noted that agent transmission is protected by the agent transport protocol (Guan & Yang, 2002), thereby protecting the system from being compromised by malicious hosts.

The key seed negotiation process is based on the Diffie-Hellman (DH) key exchange protocol (Schneier, 1996) with a variation. The agent will first generate a private DH parameter *a* and its corresponding public parameter *x*. The value *x*, together with the ID of the destination host, will be encrypted using a communication session key and sent to the agent butler.

The agent butler will decrypt the message using the same communication session key (to be discussed later). It too will generate its own DH private parameter *b* and its corresponding public parameter *y*. With the private parameter *b* and the public parameter *x* from the agent, the butler can derive the new key seed and use it for communications with the agent in the new host. Instead of sending the public parameter *y* to the agent as in normal DH key exchange, the agent butler will encrypt the value *y*, host ID, agent ID, and current timestamp with the destination host's public key to get message *M*. Message *M* will be sent to the agent after encrypting with the communication session key.

$$M = E(y + \text{host ID} + \text{agent ID} + \text{timestamp}, H_{\text{pubKey}})$$

At the same time, the agent butler updates the agent's itinerary and stores the information locally. This effectively protects the agent's actual itinerary against any hacking attempts related to itinerary, thereby protecting against the data deletion attack.

When the agent receives the double-encrypted DH public parameter *y*, it can decrypt with the communication session key. Since the decrypted result *M* is parameter *y* and some other information encrypted with the destination host's public key, the current host will not be able to find out the value of *y* and thus find out the new key seed to be used when the

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

agent reaches the destination host. It should be noted that this does not prevent the host from replacing *M* with its own version *M'* with the same host ID, agent ID, and timestamp, but different *y*. The inclusion of host ID, agent ID inside *M* can render such attack useless against SADIS. A detailed discussion on this attack can be found in the security analysis section of this article.

Subsequently, the agent will store *M* into its data segment and requests the current host to send itself to the destination host using the agent transport protocol (Guan & Yang, 2002).

Upon arriving at the destination host, the agent will be activated. Before it resumes normal operation, the agent will request the new host to decrypt message *M*. If the host is the right destination host, it will be able to use the private key to decrypt message *M* and thus obtain the DH public parameter *y*. As a result, the decryption of message *M* not only completes the key seed negotiation process, but also serves as a means to authenticate the destination host. Once the message *M* is decrypted, the host will verify that the agent ID in the decrypted message matches the incoming agent, and the host ID in the decrypted message matches that of the current host. In this way, the host can ensure that it is decrypting for a legitimate agent instead of some bogus agent. If the IDs in the decrypted messages match, the decrypted value of *y* is returned to the agent.

With the plain value of *y*, the agent can derive the key seed by using its previously generated private parameter *a*. With the new key seed derived, the key seed negotiation process is completed. The agent can resume normal operation in the new host.

Whenever the agent or the butler needs to communicate with each other, the sender will first derive a communication session key using the key seed and use this communication session key to encrypt the message. The receiver can make use of the same formula to derive the communication session key from the same key seed to decrypt the message.

The communication session key K_{CSK} is derived using the formula below:

$$K_{CSK} = \text{Hash}(\text{key\_seed} + \text{host ID} + \text{seqNo})$$

The sequence number is a running number that starts with 1 for each agent roaming session, and is reset to 1 whenever the agent reaches a new host. Each message communicated will therefore be encrypted using a different key. As this means that the butler and agent will not be able to communicate if messages are lost without detection, SADIS makes use of TCP/IP as a communication mechanism. Once the communication is re-established after a send failure, the sender will resend the previous message (encrypted using the same communication session key). The agent and the butler can therefore synchronize on communication session key calculations.

The agent encrypts host information with a data encryption key K_{DEK}. The data encryption key is derived as follows:

$$K_{DEK} = \text{Hash}(\text{key\_seed} + \text{hostID})$$

The details on encryption will be discussed in the next section.

Data Integrity Protection Protocol

The key seed negotiation protocol lays the necessary foundation for integrity protection by establishing a session-based key seed between the agent and its butler. Digital certificates also help protect the agent data integrity.

Our data integrity protection protocol comprises two parts: chained signature generation and data integrity verification. Chained signature generation is performed before the agent leaves the current host. The agent gathers data provided by the current host d_i and construct D_i as follows:

$$D_i = E(d_i + ID_{host} + ID_{agent} + \text{timestamp}, k_{DEK})$$

or

$$D_i = d_i + ID_{host} + ID_{agent} + \text{timestamp}$$

The inclusion of host ID, agent ID, and timestamp is to protect the data from possible replay attack, especially when the information is not encrypted with the data encryption key, thereby creating an unambiguous memorandum between the agent and the host. The construction of D_i also gives the flexibility to encrypt the data or keep it plain. After constructing D_i, the agent will request the host to perform a signature on the following:

$$c_i = \text{Sig}(D_i + c_{i-1} + ID_{host} + ID_{agent} + \text{timestamp}, k_{priv})$$

where c_0 is the digital signature on the agent code by its butler.

One design focus of SADIS is not only to detect data integrity compromise, but more importantly to identify malicious hosts. To achieve malicious host identification, it is an obligation for all hosts to verify the incoming agent's data integrity before activating the agent for execution. In the event of data integrity verification failure, the previous host will be identified as the malicious host.

Data integrity verification includes the verification of all the previous signatures. The verification of signature c_0 ensures agent code integrity; the verification of c_i ensures data provided by host h_i is intact. If any signature failed the verification, the agent is considered compromised.

While the process to verify all data integrity may seem to incur too much overhead and also seem somewhat redundant (e.g., why do we need to verify the integrity of d_1 in h_3 when host h_2 already verifies that), it is necessary to ensure the robustness of the protocol and to support the function of malicious host identification. Although the agent butler can eventually detect such data integrity compromise (since the agent butler must verify all signatures), there is no way to establish the identity of malicious host(s).

Security Analysis

To analyze the effectiveness and reliability of SADIS, a detailed security analysis is performed subjecting SADIS to a variety of attacks. Based on the attack targets, the various attacks to SADIS can be classified into data attack, key attack, signature attack, itinerary attack, and composite attack. Composite attack refers to attacks that are combinations of two or more of the above-mentioned attacks. The security analysis will be organized according to the above classifications.

Data Attack

Data attack refers to any attempt that aims to compromise the data carried by an agent. Compromise can be in the form of data modification, deletion, or insertion.

Considering the data modification scenario, let us assume that the data targeted is D_i provided by host i; since the agent itinerary is protected by the butler and cannot be changed, only host i can produce a valid signature if the data were to be modified. However, even if the malicious party (or even host i itself) can produce a valid signature c_i' corresponding to D_i', since c_i is chained to the signature of the next host c_{i+1}, signature verification for host $(i+1)$ will fail. Therefore, in order to perform a successful data modification attack, the malicious host must be able to forge the signatures for all hosts in the itinerary since host i. As the only way to achieve this is to obtain the private keys of all the following hosts, data modification attack is extremely difficult under SADIS.

A number of the existing data integrity protocols suffer from data deletion attack. After analyzing the root cause of the vulnerabilities, it is realized that it is extremely important to protect the agent's itinerary. If the agent's itinerary is closely guarded by the butler, any data deletion will result in modification to the agent's itinerary and thus be detected.

Key Attack

Besides direct attack on data integrity, a malicious host may attempt to attack the various keys in order to compromise data integrity. There are three different types of keys in SADIS. They are session-based key seed, communication session key, and data encryption key.

In SADIS, the key seed is kept by the agent and the butler separately. Attacks to the key seed can only target at the key seed negotiation protocol. As all communication in key seed negotiation is protected by the communication session key, we can safely rule out the possibility of any third-party malicious attempts to break the protocol. We can focus on the scenario where the current host attempts to break the key exchange to obtain the key seed to be used in the subsequent host.

Firstly, as the DH public parameter is encrypted using the destination host's public key, the current host will not, without manipulation, be able to complete DH key exchange to find out the new key seed. Without the private key from the destination host, no one can obtain y to complete the key exchange. Furthermore, as the encrypted message contains the agent ID and destination host ID, the current host will not be able to send a bogus agent carrying this encrypted y to the destination host for decryption.

If the current host attempts to manipulate any one or both of these parameters, it is able to manipulate the key seed derived when the agent reaches the destination host. However, the change in key seed will be immediately detected when the agent communicates with the butler or vice versa. In order to perform a successful attack, the current host must therefore be able to obtain the key seed in the butler so that it can intercept and replace the message communicated between the butler and the agent. Unfortunately, as illustrated earlier, there is no way the current host can find out the value of DH public parameter from butler y. Thus, the key seed will not be compromised.

Besides key seed, SADIS makes use of communication session key and data encryption key in the protocol. These two keys are directly derived from the session-based key seed using a hash function. As far as any third-party host is concerned, attack to communication session key or data encryption key is equivalent to attacking the encryption key given only the cipher text. Even in the extreme case when such a key is compromised, the loss is limited to the message it encrypts.

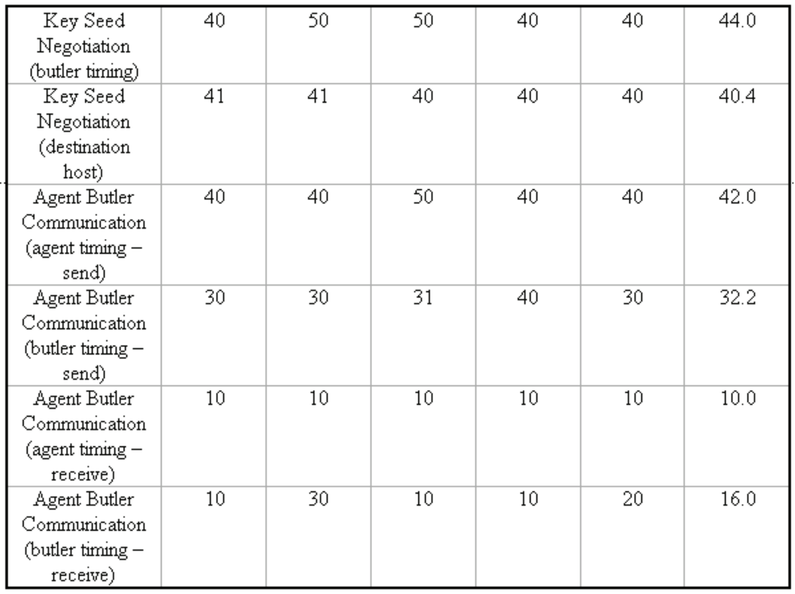
Signature Attack

Usually a malicious host would need to forge digital signature when it attempts to compromise data integrity. If data integrity is not compromised, there is no need to attack the chained signature at all.

Itinerary Attack

If the agent itinerary is not carefully protected, it may lead to compromise to data integrity, especially in the case of data deletion as illustrated earlier in the section. In SADIS, as the agent updates the butler of its next destination host as part of the key seed negotiation protocol, there is no additional

Table 1. SADIS time efficiency (time taken for operations in milliseconds): Performance without SADIS



Key Seed Negotiation (butler timing)	40	50	50	40	40	44.0
Key Seed Negotiation (destination host)	41	41	40	40	40	40.4
Agent Butler Communication (agent timing – send)	40	40	50	40	40	42.0
Agent Butler Communication (butler timing – send)	30	30	31	40	30	32.2
Agent Butler Communication (agent timing – receive)	10	10	10	10	10	10.0
Agent Butler Communication (butler timing – receive)	10	30	10	10	20	16.0

overhead related to the itinerary protection mechanism. Therefore, there is no way a malicious host can perform any attack on the itinerary (except, of course, if it breaks into the agent butler).

Composite Attack

At times, in order to perform a successful attack, more than one area is targeted simultaneously. In addition to attacks with specific targets, there are certain general hacking techniques such as man-in-the-middle attack, replay attack. The design of SADIS employs a mechanism to protect the protocol against these hacking techniques. Through the use of communication session key, man-in-the-middle attack can be avoided. On the other hand, the use of sequence number in communication session key generation effectively protects the protocol from replay attack by a third-party host. In addition, the inclusion of host ID, agent ID, and timestamp during the key seed negotiation process prevents the current host from performing a replay attack with the next destination host.

Lastly, the design of SADIS does not have dependency on any specific encryption/hashing algorithm. In an unlikely scenario when one algorithm is broken, SADIS can always switch to a stronger algorithm.

Implementation

In order to verify the design of SADIS and assess its applicability, a prototype of SADIS is developed. The prototyping language is chosen to be Java, because of its platform-independent feature.

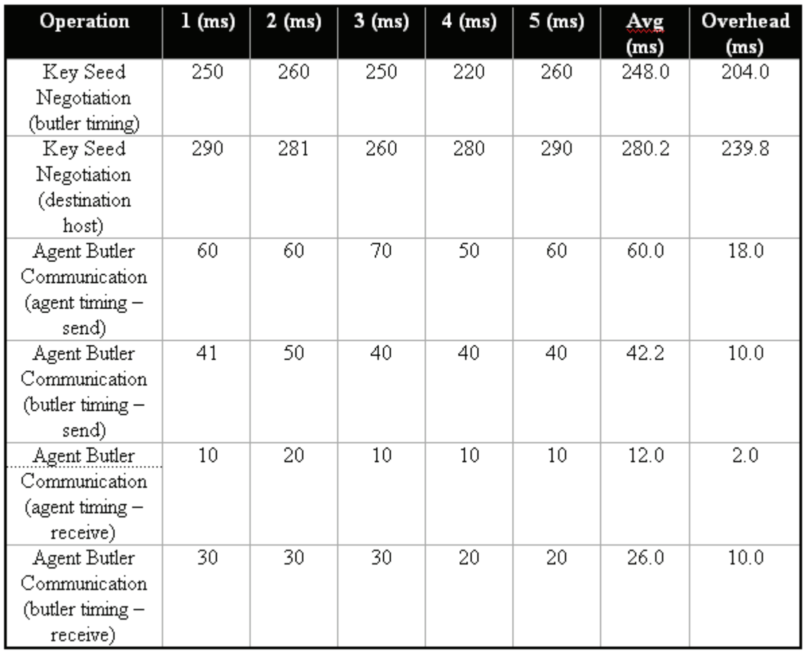
Just like any other security mechanism, there is certain overhead associated with SADIS. The overhead is incurred as additional time required for processing as well as additional data carried by the agent. To assess the efficiency of SADIS, a study is performed on the prototype.

The result of this experimental study on SADIS is broken down based on functionality and is shown in Tables 1 and 2. It can be seen that the bulk of the overhead is incurred during key seed negotiation where the key exchange protocol and the public key operation is performed. Despite the relatively high overhead, this will not impact the overall performance of SADIS significantly because the frequency of agent roaming is low compared to the frequency of some other agent operations (such as agent to butler communication). As a result, the overhead incurred at this stage is 'one-time' in nature.

Other than in the key seed negotiation, the time overhead incurred elsewhere in the protocol is negligible.

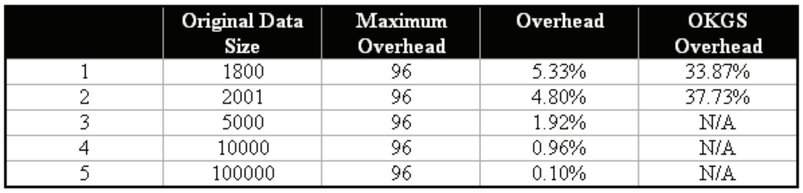
Other than overhead in terms of processing time, there is certain overhead to the data size as well. SADIS is designed to produce almost fixed data overhead regardless of the data size. SADIS therefore tends to be more efficient when actual data size is higher. This ability to limit the size of overhead data regardless of actual data size is an improvement in efficiency over existing work. The last and most significant overhead is the digital signature created by the host. The overhead of digital signature is a fixed length of 64 bytes. Altogether, SADIS has a maximum data overhead of 96 bytes.

Table 2. SADIS time efficiency (time taken for operations in milliseconds): Performance comparison with SADIS



Operation	1 (ms)	2 (ms)	3 (ms)	4 (ms)	5 (ms)	Avg (ms)	Overhead (ms)
Key Seed Negotiation (butler timing)	250	260	250	220	260	248.0	204.0
Key Seed Negotiation (destination host)	290	281	260	280	290	280.2	239.8
Agent Butler Communication (agent timing – send)	60	60	70	50	60	60.0	18.0
Agent Butler Communication (butler timing – send)	41	50	40	40	40	42.2	10.0
Agent Butler Communication (agent timing – receive)	10	20	10	10	10	12.0	2.0
Agent Butler Communication (butler timing – receive)	30	30	30	20	20	26.0	10.0

Table 3. SADIS data overhead



	Original Data Size	Maximum Overhead	Overhead	OKGS Overhead
1	1800	96	5.33%	33.87%
2	2001	96	4.80%	37.73%
3	5000	96	1.92%	N/A
4	10000	96	0.96%	N/A
5	100000	96	0.10%	N/A

As the statistics shows, SADIS is optimized to improve both time efficiency and data efficiency compared with related work in the literature. The feasibility and practicality of SADIS is thus demonstrated through the prototype.

IMPACT OF SADIS

Various techniques have been developed to protect agent integrities (Borselius, 2002), based on trusted hardware, trusted host, and conventional contractual agreements. SADIS addresses the problem of data integrity protection via a combination of techniques such as execution tracing, encrypted payload, environmental key generation, and undetachable signature. The security of SADIS is completely based on its own merits without making any assumption about the integrity of external hosts. SADIS also makes use of a negotiated key seed to generate data encryption key. Therefore, no random value needs to be encrypted and stored with the agent. With SADIS, the data and the communication keys undergo one-time encryption. Thus, even if some of the keys are compromised, the key seed will still remain secret.

CONCLUSION

In this article, a new data integrity protection protocol, SADIS, has been proposed. Besides being secure against a variety of attacks and robust against vulnerabilities of related work in the literature, the research of SADIS includes the objective of efficiency. Unlike some existing literature, the data integrity protection protocol aims not only to detect data integrity compromise, but more importantly to identify the malicious host. With security, efficiency, and effectiveness as its main design focuses, SADIS works with other security mechanisms to provide mobile agents with a secure platform.

REFERENCES

Borselius, N. (2002). Mobile agent security. *Electronics & Communication Engineering Journal, 14*(5), 211-218.

Chionh, H. B., Guan, S.-U., & Yang, Y. (2001). Ensuring the protection of mobile agent integrity: The design of an agent monitoring protocol. *Proceedings of the IASTED International Conference on Advances in Communications* (pp. 96-99).

Guan, S.-U., & Yang, Y. (2004). Secure agent data integrity shield. *Electronic Commerce and Research Applications, 3*(3), 311-326.

Guan, S.-U., & Yang, Y. (2002). SAFE: Secure-Roaming Agents for E-commerce. *Computers & Industrial Engineering Journal, 42,* 481-493.

Guan, S.-U., Zhu, F., & Maung, M.-T. (2004). A factory-based approach to support e-commerce agent fabrication. *Electronic Commerce and Research Applications, 3*(1), 39-53.

Guan, S.-U., & Zhu, F. (2004). Ontology acquisition and exchange of evolutionary product-brokering agents. *Journal of Research and Practice in Information Technology, 36*(1) 35-46.

Johansen, D., Lauvset, K. J., Renesse. R., Schneider, F. B., Sudmann, N. P., & Jacobsen, K. (2002). A Tacoma retrospective. *Software—Practice and Experience,* 605-619.

Schneier, B. (1996). *Applied cryptography: Protocols, algorithms, and source code in C* (2nd ed.). New York: John Wiley & Sons.

Tianhan, W., Guan, S.-U., & Chan, T. K. (2002). Integrity protection for code-on-demand mobile agents in e-commerce. *Journal of Systems and Software, 60*(3), 211-221.

Yang, Y., & Guan, S.-U. (2000). Intelligent mobile agents for e-commerce: Security issues and agent transport. In *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.

Zhu, F., Guan, S.-U., Yang, Y., & Ko, C.C. (2000). SAFER e-commerce: Secure agent fabrication, evolution and roaming for e-commerce. In *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.

KEY TERMS

Agent: A piece of software which acts to accomplish tasks on behalf of its user.

Cryptography: The art of protecting information by transforming it (*encrypting* it) into an unreadable format, called cipher text. Only those who possess a secret *key* can decipher (or *decrypt*) the message into plain text.

Digital Certificate: Certificate that uses a digital signature to bind together a public key with an identity—information such as the name of a person or an organization, their address, and so forth. The certificate can be used to verify that a public key belongs to an individual.

Electronic Commerce (E-Commerce): Consists primarily of the distributing, buying, selling, marketing, and servicing of products or services over electronic systems such as the Internet and other computer networks.

Flexibility: The ease with which a system or component can be modified for use in applications or environments other than those for which it was specifically designed.

Protocol: A convention or standard that controls or enables the connection, communication, and data transfer between two computing endpoints. Protocols may be implemented by hardware, software, or a combination of the two. At the lowest level, a protocol defines a hardware connection.

Security: The effort to create a secure computing platform, designed so that agents (users or programs) can only perform actions that have been allowed.

Secure Group Communications in Wireless Networks

Yiling Wang
Monash University, Australia

Phu Dung Le
Monash University, Australia

INTRODUCTION

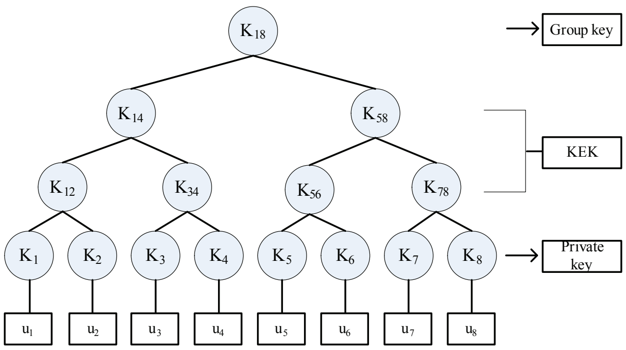
After a decade of exponential growth, wireless technologies have profoundly impacted people's lifestyles. Wireless networks provide users with greater flexibility and benefit by anytime and anywhere services (Pahlavan & Krishnamurthy, 2002). At the same time, rapid developments in multicast have led to the emergence of many multicast applications, such as stock quoting, multimedia conferencing and network gaming. Therefore it is reasonable to believe that the integration of wireless and multicast will benefit mobile users. Before the customers can enjoy the efficiency and convenience of wireless multicasts, access control must be employed to guarantee that only legitimate users can utilize the multicast services.

BACKGROUND

Access control can be achieved by employing a cryptographic key shared by all group members, which is applied to encrypt the multicast contents. Many group key management approaches (Sherman & McGrew, 2002; Perrig, Song, & Tygar, 2001; Amir, Kim, NitaRotaru, Schultz, Stanton, & Tsudik, 2004; Harney & Muckenhirn, 1997; Steiner, Tsudik, & Waidner, 1996; Mittra, 1997; Banerjee & Bhattacharjee, 2002; Kostas, Kiwior, Rajappan, & Dalal, 2003) have been proposed and most of them are directed toward the wired network. Although the research work done in the wired environment can be applied in wireless networks, the efficiency and security are not the same as that in wired networks. The reason results from not only the limitations of wireless networks, such as high communication error rate and limited bandwidth, but also the properties of light-weight mobile devices, such as limited computational power, insufficient power supply, great mobility and storage limitation.

In order to reduce the communication, computation and storage costs, hierarchical structure (tree structure) is widely applied in the group key management approaches. These schemes utilize all the keys, that is, group key and supporting keys, to construct a balanced key tree.

Figure 1. A typical key management tree



Each node of the tree holds a key. The root node of the tree represents the group key. Each leaf node corresponds to a group member, and possesses a private key associated with the member. The intermediate nodes hold key encryption keys (KEK), which are the auxiliary keys used for the distribution of the group key and other KEKs. Each member needs to store a set of keys in the path from its node to the root of the tree. When a member joins or leaves, the key distributor center (KDC) generates a set of new keys from the leaf node associated with the member to the root, and multicasts this set of keys to all the other group members. For example, as shown in Figure 1, for user 3 joining or leaving, key $k_{34,}$ $k_{14,}$ and k_{18} need to be updated.

For user 3 joining:

$$\mathrm{KDC} \rightarrow u_3\!:\{k_{34}, k_{14}, k_{18}\}_{k_3}$$

$$\mathrm{KDC} \rightarrow u_4\!:\{k_{34}, k_{14}, k_{18}\}_{k_4}$$

$$\mathrm{KDC} \Rightarrow (u_1,u_2)\!:\{k_{14}, k_{18}\}_{k_{12}}$$

$$\mathrm{KDC} \Rightarrow (u_5,u_6,u_7,u_8)\!:\{k_{18}\}_{k_{58}}$$

For user3 leaving:

$$\mathrm{KDC} \rightarrow u_4\!:\{k_{34}, k_{14}, k_{18}\}_{k_4}$$

$$\mathrm{KDC} \Rightarrow (u_1,u_2)\!:\{k_{14}, k_{18}\}_{k_{12}}$$

$$\mathrm{KDC} \Rightarrow (u_5,u_6,u_7,u_8)\!:\{k_{18}\}_{k_{58}}$$

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

GROUP KEY MANAGEMENT ALGORITHM

Group key management algorithm is core part of multicast security. It maintains the logical key structure and performs the procedures to assign, distribute and update the group key and other KEKs.

Notation

In this section, we depict the notations that we will use in the following sections:

Þ u: user

Þ bs: base station

Þ n: the number of members in the subgroup

Þ n_c: the number of the cluster in the subgroup

Þ n_s: the number of subgroups

Þ m: the number of users in the cluster

Þ j: the number of multiple subgroups which user joins and leaves simultaneously

Þ α: degree of the balance tree

Þ d: the height of the balance tree ($d = \log_\alpha n$)

Þ k: the encyption key

Þ BS = $\{s_1, s_2, s_3, ..., s_n\}$: the set of the base stations

Þ {x}k: message x encrypted by the key k

Þ A → B: {x}: A sends message x to B via unicast

Þ A => B: {x}: A sends message x to B via broadcast or multicast

The Proposed Logical Key Structure

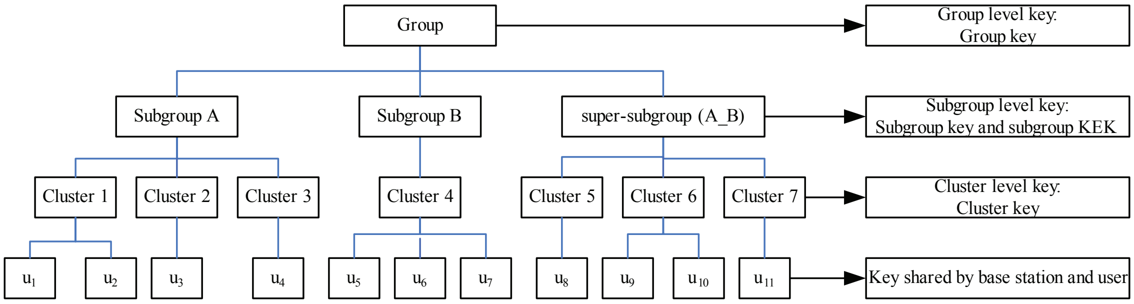
Generally, a large multicast group is comprised of several smaller subgroups. Each member not only joins the group communications but also participates in one or some subgroup communications. Nevertheless the current proposed key tree cannot reflect such a group organization structure. Meanwhile a separate key tree must be constructed for each subgroup. To improve the performance of group key management under such scenario, we propose a new logical group keying structure shown in Figure 2.

From Figure 2, we can see that the proposed structure is a multi-tier model. The root node instigates the group communication session, and holds the group key. Each subgroup represents a multicast session and is associated with two keys: subgroup key and subgroup KEK (key encryption key). Users in the subgroup are divided into several clusters. Each cluster has its own cluster key to distribute other keys in the cluster. In the user level, each user shares a secret key with base station.

It is a common scenario that a user subscribes multiple subgroups simultaneously. To improve the efficiency, we introduce a new concept: super-subgroup, which is a virtual container to accommodate users who participate in multiple subgroups simultaneously. Each super-subgroup is a combination of several subgroups. For example, as shown in Figure 2, super-subgroup (A_B) is a super-subgroup combining subgroup A and B, where users 9 to 11 participate in the subgroup A and B concurrently. Super-subgroup can be a combination of any subgroups. According to the theory of permutation and combination, for a group having n_s subgroups, the total number of super-subgroups is

$$\sum_{i=2}^{ns} C_n^i \, .$$

Figure 2. Logical key management structure



The Proposed Group Key Management Algorithm

There are three main operations in the wireless group key management (multiple subgroups): member join, member leaving and hand-off. The rekeying procedures of these operations occur independently in each wireless cell. We now illustrate our algorithm in each of these operations in the following subsections.

Member Join

Join is a procedure that is invoked by a user who wants to become a member of a group. Backward secrecy should be achieved to prevent the new member from accessing the previous group communication contents. In the proposal, join is comprised of three steps: registration, key grant and key update. In the registration step, a user submits a join request to the local base station, and the base station forwards this request to upper layer area key controller (AKC) for authentication.

u → bs: {JOIN_REQUEST}

bs → AKC: {JOIN_REQUEST}

In the second step, after AKC verifies the user, AKC assigns the user into a cluster and sends the relevant keys to the user.

$\text{AKC} \rightarrow \text{bs}: \{k_{\text{cluster,}}\ k_{\text{subgroup,}}\ k_{\text{subgroup\_kek,}}\ k_{\text{group}}\}_{\text{k (AKC\_user)}}$

$\text{bs} \rightarrow \text{u}: \{k_{\text{cluster,}}\ k_{\text{subgroup,}}\ k_{\text{subgroup\_kek}},\ k_{\text{group}}\}_{\text{k (AKC\_user)}}$

Finally, AKC invokes a join-key-update procedure to update the affected cluster, subgroup and group key. This updating can be achieved without multicasting rekeying message (Waldvogel, Caronni, Sun, Weiler, & Plattner, 1999). AKC informs the sender to use the new keys to encrypt the data traffic; the members will perceive the key change in ordinary data packets and update their keys locally by passing through a one-way hash function.

AKC → sender: {USING NEW DATA ENCRYPTION KEY}

When a user wishes to join multiple subgroups at the same time, the procedure is the same as mentioned above. The user is assigned into a suitable super-subgroup.

Member Leaving

Leaving is invoked by a user who wants to quit a group or the AKC evicts the user from the group communications. Forward secrecy must be achieved to prevent the departing user obtaining the forthcoming group communications. There are two steps in this operation: cancellation and key update. In the cancellation, user submits a leaving request to AKC:

u → bs: {LEAVE_REQUEST}

bs → AKC: {LEAVE_REQUEST}

In the rekeying stage, according to the "bottom to top" principle, rekeying procedure starts from the affected cluster where the departure took place then to the remaining clusters in the affected subgroup, and finally in the entire group. First, AKC generates new keys, and constructs a single message containing all new keys by the *group-oriented* approach (Wong, Gouda, & Lam 2000).

$\text{AKC} \Rightarrow \text{BS}: \{\text{new } k_{\text{cluster}},\ k_{\text{subgroup\_KEK}},\ k_{\text{subgroup,}}\ k_{\text{group}}\}$

Each base station searches for the affected cluster members in its cell and delivers the new keys to them. Base station employs *user-oriented rekeying* approach (Wong et al., 2000), that is, for each user the base station constructs a rekeying message that contains precisely the new keys needed by the user.

$\text{bs } u_i: \{\text{new } k_{\text{cluster}},\ k_{\text{subgroup\_KEK}},\ k_{\text{subgroup,}}\ k_{\text{group}}\}_{\text{k(bs\_user) i(i=1, 2...m)}}$

To update keys for the remaining clusters in the affected subgroup, the base station multicasts the rekeying message to the clusters members. The new keys are encrypted by the old group key:

$\text{bs} \Rightarrow (\text{cluster})_i: \{\text{new } k_{\text{subgroup\_KEK}},\ k_{\text{subgroup,}}\ k_{\text{group}}\}_{\text{k(cluster)i(i=1, 2\dots } n_c)}$

Finally, the base station updates the group key for the remaining subgroups in the group.

$\text{bs} \Rightarrow (\text{subgroup})_i: \{\text{new } k_{\text{group}}\}_{\text{k (subgroup\_KEK) i (i = 1, 2, \dots , } n_s)}$

If a multi-subgroup user leaves, the procedure is same.

$\text{bs} \Rightarrow (\text{subgroup})_i: \{\text{new } k_{\text{subgroup,}}\ k_{\text{group}}\}_{\text{k (subgroup\_KEK) i (i = 1, 2, \dots , j)}}$

Handoff

Handoff is a unique operation in the wireless group communications. Several approaches (DeCleene, Dondeti, Griffin, Hardjono, Kiwior, Kurose, Towsley, Vasudevan, & Zhang, 2001; Sun, Trappe, & Liu 2002) have been proposed to address the group key management during the handoff.

We propose a simple and efficient handoff management scheme, which contains three steps: handoff registration, authentication and switching. We assume that the mobile devices can detect signals of the two adjacent base stations on the border of the wireless cell (Chen & Chao, 2004). In the registration step, user switches to the new base station and sends a handoff join message. Then user switches back to the old base station quickly.

u → bs_{new}: {HANDOFF_JOIN}

In the authentication stage, new base station contacts the old base station to verify the user:

bs_{new} → bs_{old}: {AUTHENTICATION_REQUEST}
bs_{old} → bs_{new}: {AUTHENTICATION_REPLY}

In the last step, after a predefined time, user switches to the new base station to finish the handoff. Because the group keying structure is identical in the whole domain, users can move from one cell to another without invoking rekeying operation.

PERFORMANCE DISCUSSION

In measuring the performance of a group key management system, many parameters need to be taken into consideration (Moyer, Rao, & Rohatgi, 1999; Rafaeli & Hutchison, 2003).

However, efficiency is one of the most significant criteria. Communication, computation and storage overheads related to rekeying are the major efficiency measures for a key management algorithm. We take the popular tree-based key management scheme: logical key hierarchy (LKH) (Wong et al., 2000; Wallner, Harden, & Agee, 1999) as the benchmark in our performance analysis.

Communication Efficiency

In order to achieve the best result, the communication during the rekeying and handoff combines the unicast and multicast. Communication overhead is recorded in *big-O* notation as a measure of the number of rekeying messages transmitted per operation. We evaluate the communication overhead for the following three operations: join, leave and handoff.

Join

As described above, when a user joins a group having n_s subgroups, there is no need to multicast rekeying message. The system only needs to inform the sender to apply the new keys. So the communication cost in this situation is zero. According to the LKH algorithm, the scheme needs to construct separate n_s trees to present all the subgroups. The communication cost of join is d (Wong et al., 2000; Wallner, Harden, & Agee, 1999). When a single subgroup join occurs, there is a join operation and group key updating for the rest (n_s-1) subgroups. So the cost of rekeying communication for LKH is O(d+n_s).

When a user wants to join multiple subgroups simultaneously, with the concept of super-subgroup, the cost of communication in the proposal is still zero. As for the LKH algorithm, multiple trees are affected by the join operation and n_s–j subgroups need to update their group key. So in such scenario, the communication overhead of LKH is the summation of cost of multiple join operation and n_s–j group key updating, that is,

$$O\left(\sum_{i=1}^{j} d_i + n_s - j\right).$$

Leaving

When a single subgroup member leaves, the proposed algorithm needs to rekey four keys: cluster, subgroup, subgroup KEK and group key. The overhead of rekeying communication is shared by AKC and base station network. AKC is responsible for the generation of new keys. By using the group-oriented approach, all the new keys are contained in a single rekeying message, so the communication cost of AKC is O(1). Base station is in charge of the key distribution in its cell. There are two scenarios in the cell: the leaving user present in the cell or the leaving user not in the cell. In the latter situation, the base station can use multicast for rekeying, the rekeying communication is just one or two messages. Therefore we focus our attention on the former scenario. According to our proposal, first, base station sends the rekeying message to the affected cluster users via unicast, so the cost is $m-1$. Then, base station multicasts the rekeying message to the members of the remaining clusters in the affected subgroup. The overhead of this communication is n_c-1. Finally, base station needs to update the group key for the rest of subgroups in the group and the cost is n_s-1. Thus the total cost of rekeying communication of base station is O(m +n_c+ n_s). As for the LKH, the leaving cost is $(\alpha - 1)d$ (Wong et al., 2000; Wallner, Harden, & Agee 1999). When single subgroup departure happens, the communication cost of LKH is one leaving cost plus group key updating for (n_s-1) subgroups. Therefore the overhead of LKH for single subgroup leaving is O $((\alpha - 1)d + n_s)$.

When a member quits multiple subgroups simultaneously, multiple subgroups are affected. Under such scenario, AKC's communication cost is still O(1), because all the new keys are still in one message. By using the subgroup KEK, base station has the same cost as that of single subgroup leav-

Table 1. The communication cost comparison of our proposal and LKH

	join		Leaving		handoff
	Single	Multi-subgroup	Single	Multi-subgroup	
our scheme	zero	zero	AKC: O(1)	AKC: O(1)	zero
			BS: O(m +n_c+ n_s)	BS: O(m +n_c+ n_s)	
LKH	O(d+n_s)	$O(\sum_{i=1}^{j} d_i + n_s - j)$	$O((\alpha - 1)d + n_s)$	$O(\sum_{i=1}^{j} (\alpha - 1)d_i + n_s - j)$	O(d)

Note: n_c = the number of cluster of the subgroup; $n_{s:}$ = the number of the subgroups; M = the size of the cluster; α = the degree of the balance tree; j = the number of multiple subgroups which user joins or leaves simultaneously; d = the height of the tree

ing: O(m +n_c+ n_s). As we described above, LKH algorithm needs multiple leaving operations. Additionally, the group key of (n_s–j) subgroup needs to be updated. So the overhead of communication of LKH is:

$$O(\sum_{i=1}^{j} (\alpha - 1)d_i + n_s - j).$$

Handoff

Because of the identical logical keying structure in all the areas, users can freely move from one cell to another without invoking rekeying procedure. As for the LKH algorithm, some handoff procedures at least need a join operation (DeCleene et al., 2001). So the rekeying communication cost is O(d), d is the height of the balance tree.

Summary

We tabulate the efficiency evaluation in Table 1. From the table, we can see that our proposal has advantages over the LKH, especially in the handoff and multi-subgroup join and leaving.

Computation Efficiency

Computation efficiency is to measure the cost of computation during the rekeying process in the wireless devices. Because of the computation limitation of the wireless device, it can not afford the expensive computation function, such as exponentiation, PKI calculation and so on. In our proposal, there are only two kinds of computations performed in the wireless devices: one-way hash function and symmetric encryption/decryption. In reference to the current mobile technology, these calculations were confirmed that they could be operated in a fast and efficient manner in the wireless devices. So the proposed system can achieve good computation efficiency which makes it suitable for the wireless network environment.

Storage Efficiency

The storage efficiency is to measure the number of keys stored in AKC and mobile devices. We compare the key storage cost of our proposal with that of LKH (Wong et al., 2000; Wallner, Harden, & Agee 1999) in Table 2.

From the table, we can see that, on the server side, the keys stored in LKH algorithm are linearly proportional to the number of group users. In the proposal, the key storage cost is proportional to the number of clusters, which is much less than the number of users. On the user side, there are only 4 keys stored for a single subgroup membership and j+3 keys for the multi-subgroup memberships in our proposal. For the LKH, in the same scenarios, users need to store d + 1 and

$$\sum_{i=1}^{j} d_i + j + 1$$

keys in single and multiple subgroup scenario respectively.

Table 2. The key storage cost comparison of our proposal and LKH

	Our proposal	LKH
KDC/AKC	$1 + \sum_{i=1}^{n_s} (n_c)_i + n_s$	$\sum_{i=1}^{n_s} \frac{\alpha n_i - 1}{\alpha - 1} + 1$
Users (single subgroup)	4	d + 1
Users (multi-subgroup)	j + 3	$\sum_{i=1}^{j} d_i + j + 1$

Note: N = the number of subgroup members; n_c = the number of cluster in the subgroup; $n_{s:}$ = the number of the subgroups; α = the degree of the balance tree; j = the number of multiple affected subgroups; d = the height of the tree

FUTURE TRENDS

Along with the fast development of wireless communications and fast capacities improvement on mobile devices, more and more multicast group applications and services will be emerging on wireless networks. The future research will focus on the efficiency and security of group key management system. Additionally, new architecture and framework will be proposed to address the wireless multicast security.

CONCLUSION

Here, we present a new group key management solution to secure multicast communications in wireless networks. This proposed solution has distributed two-tier architecture and clustered hierarchical keying structure based on the group organization chart. The group key management system can perform the multi-subgroup access control in an efficient way. Compared with the existing tree-based key management scheme, this proposed scheme can significantly reduce the overhead associated with communication, computation and storage.

REFERENCES

Amir, Y., Kim, Y., NitaRotaru, C., Schultz, J. L., Stanton, J., & Tsudik, G. (2004). Secure group communication using robust contributory key agreement*. IEEE Transactions on Parallel and Distributed Systems*, *15*(5), 468-480

Banerjee, S., & Bhattacharjee, B. (2002). Scalable secure group communication over IP multicast. *IEEE Journal on Selected Areas in Communications*, *20*(8), 1511-1527

Chen, J., & Chao, T. (2004). *IP-based next-generation wireless networks*. NJ: John Wiley & Sons.

DeCleene, B., Dondeti, L. D., Griffin, S., Hardjono, T., Kiwior, D., Kurose, J., et al. (2001). Secure group communications for wireless networks. Military Communications Conference. Communications for Network-Centric Operations: Creating the Information Force. *IEEE*, *1*, 113-117.

Harney, H., & Muckenhirn, C. (1997). *Group key management protocol (KGMP) architecture*. RFC 2094.

Kostas, T., Kiwior, D., Rajappan, G., & Dalal, M. (2003). Key management for secure multicast group communication in mobile networks. In *Proceedings of DARPA Information Survivability Conference and Exposition* (Vol. 2, pp. 41-43).

Mittra, S. (1997). Iolus: A framework for scalable secure multicasting. *Processing of the ACM SIGCOMM*, *27*(4), 277-288

Moyer, M. J., Rao, J. R., & Rohatgi, P. (1999). A survey of security issues in multicast communications. *IEEE Network*, *13*, 12-23

Pahlavan, K., & Krishnamurthy, K. (2002). *Principles of wireless networks: A unified approach*. NJ: Pearson Education, Inc.

Perrig, A., Song, D., & Tygar, D. (2001). ELK, A new protocol for efficient large-group key distribution. In *Proceedings of IEEE Symposium on Security and Privacy* (pp. 247-262).

Rafaeli, S., & Hutchison, D. (2003). A survey of key management for secure group communication. *ACM Computing Surveys*, *35*(3), 309-329.

Sherman, A. T., & McGrew, D. A. (2003). Key establishment in large dynamic group using one-way function trees. *IEEE on Software Engineering*, *29*(5), 444-458

Steiner, M., Tsudik, G., & Waidner, M. (1996). Diffie-hellman key distribution extended to group communication. In *Proceedings of the 3rd ACM Conference on Computer and Communications Security* (pp. 31-37).

Sun, Y., Trappe, M., & Liu, K. J. R. (2002). An efficient key management scheme for secure wireless multicast. In *Proceedings of IEEE International Conference on Communication* (pp. 1236-1240).

Waldvogel, M., Caronni, G., Sun, D., Weiler, D., & Plattner, B. (1999). The Versakey framework: Versatile group key management. *IEEE Journal on Selected Areas in Communications*, *17*(8), 1- 15.

Wallner, D. M., Harden, E. J., & Agee, R. C. (1999). *Key management for multicast: Issues and architecture*. RFC 2627.

KEY TERMS

Backward Secrecy: To prevent new group members from accessing previous group communications, which they may have recorded.

Forward Secrecy: To prevent departing members from decoding future group data traffic.

Handoff: In a cellular wireless network, handoff is the transition of signal for any given user from one base station to a geographically adjacent base station as the user moves around.

Key Encryption Key (KEK): This kind of keys is used to encrypt the other keys for distribution in the multicast group.

Key Management Algorithm: In the group key management system, an algorithm is applied to maintain the logical key structure held by the group members and other entities.

Logical Key Hierarchy (LKH): This algorithm is a tree structure for efficient group rekeying. Each node of the tree represents a key, with the root node representing the group key. Each leaf node represents a group member, and each member knows all the keys in its path to the root.

Multicast: Multicast is communication mechanism to delivery a single message to multiple receivers on a network. The message will be duplicated automatically by routers when multiple copies are needed.

Security Architectures of Mobile Computing

Kaj Grahn
Arcada Polytechnic, Finland

Göran Pulkkis
Arcada Polytechnic, Finland

Jonny Karlsson
Arcada Polytechnic, Finland

Dai Tran
Arcada Polytechnic, Finland

INTRODUCTION

Mobile Internet users expect the same network service quality as over a wire. Technologies, protocols, and standards supporting wired and wireless Internet are converging. Mobile devices are resource constrained due to size, power, and memory. The portability making these devices attractive also causes data exposure and network penetration risks.

Mobile devices can connect to many different wireless network types, such as cellular networks, personal area networks, wireless local area networks (WLANs), metropolitan area networks (MANs), and wide area networks (satellite-based WANs). Wireless network application examples are e-mailing, Web browsing, m-commerce, electronic payments, synchronization with a desktop computer, network monitoring/management, and reception of video/audio streams.

BACKGROUND

Major security threats for mobile computing devices are (Olzak, 2005):

- theft/loss of the device and removable memory cards,
- wireless connection vulnerabilities, and
- malicious code.

Mobile computing devices are small, portable, and thus easily lost/stolen. Most mobile platforms only include support for simple software-based password login schemes. These schemes are easily bypassed by reading information from the device without login. Memory cards are also easily removed from the device.

Mobile devices support wireless network connections such as Bluetooth and WLAN. These connections are typically by default unprotected and thus exposed to eavesdropping, identity theft, and denial-of-service attacks.

Malware has constituted a growing threat for mobile devices since the first Symbian worm (Cabir) was detected in 2004. Mobile devices can be infected via MMS, Bluetooth, infrared, WLAN, downloading, and installing from the Web. Current malware is focused on Symbian OS and Windows-based devices. Malware may result in (Olzak, 2005):

- loss of productivity,
- exploitation of software vulnerabilities to gain access to resources and data,
- destruction of information stored on a SIM (subscriber identity module) card, and
- hi-jacking of airtime resulting in increased costs.

WIRELESS SECURITY PRINCIPLES

Security Policy

Examples of rules proposed for mobile device end users are:

- I agree to make sure my device is password protected and that latest security patches are installed.
- I agree to keep a firewall/anti-virus client with latest anti-virus signatures installed, and to use a remote access VPN client, if I will connect to the corporate network.
- I agree to use the security policies recommended by the corporate security team.

Examples of rules proposed for administrators of mobile devices in corporate use are:

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- End-users get mobile network access after agreeing to the end-user rules of behavior.
- Handheld firewalls shall be configured to log security events and send alerts to *security-manager@company.com.*
- Handheld groups and Net groups shall have restricted access privileges and only to needed services.

Handheld security policies should be automated by restrictive configuration settings for handhelds, firewalls, VPNs, intrusion detection systems, and directory servers (Handheld Security, 2006).

Storage Protection

Mobile device storage protection is online integrity control of all stored program code and all data, optional confidentiality of stored user data, and protection against unauthorized tampering of stored content. Protection should include all removable storage modules used by the mobile device.

The integrity of the operating system code, the program code of installed applications, and system and user data can be verified by checksums, cyclic redundancy codes (CRCs), hashes, message authentication codes (MACs, HMACs), cryptographic signatures, and so forth. However, only hardware protection of verification keys needed by MACs, HMACs, and signatures provide strong protection against tampering attacks. Online integrity control of program and data files must be combined with online integrity control of the configuration of a mobile device for protection against malware intrusion attempts.

User data confidentiality can be granted by file encryption software. Such software also protects integrity of stored information, since successful decryption of an encrypted file is also an integrity proof.

Security Layers

Mobile computing security layers are based on the OSI (Open Systems Interconnection) Security Model. Defined security services are *authentication, access control, non-repudiation, data integrity, confidentiality*, *assurance/availability,* and *notarization/signature* (ISO/IEC 7498-1, 1994; ISO 7498-2, 1989).

Specific wireless security architecture issues include Mobile IP security features, and link-level and physical-level security protocols of wireless access technologies like WLAN, GPRS, and Bluetooth

Mobile IP security means that:

- a mobile node, which is a mobile device, has the same connectivity and security in a visited foreign network as in its home network; and
- the home network and visited foreign networks have protection against active/passive attacks.

These security goals require:

- that Mobile IP registration and location update messages have *data integrity protection, data origin authentication,* and *anti-replay protection*;
- *access control* to foreign network resources used by visiting mobile nodes; and
- that IP packet redirecting tunnels provide *data integrity protection, data origin authentication,* and *data confidentiality.*

Moreover, mobile nodes should have *location privacy* and *anonymity* (Zao et al., 1999).

Replay prevention with timestamps or nonces for all mobile IP messages is specified in Perkins and Calhoun (2000). Other mobile IP security solutions are authentication schemes and protection of data communication (Calhoun et al., 2005; Barun & Danzeisen, 2001; Hwu, Chen, & Lin, 2006).

Identification Hardware

Identification hardware contains user information and cryptographic keys used to authenticate users to mobile devices, applications, networks, and network services.

The following identification hardware types are used:

- subscriber identity module (SIM),
- public key infrastructure SIM (PKI SIM),
- universal SIM (USIM), and
- IP multimedia services identity module (ISIM).

SIM

A basic SIM card is a smartcard securely storing a key (Ki) identifying a GSM network user. A SIM card is a microcomputer executing cryptographic operations with Ki. The SIM card also stores SMS (short message service) messages, MMS (multimedia messaging system) messages, and a phonebook. The use and content of a SIM card is PIN protected (Rankl & Effing, 2003).

PKI SIM

A PKI SIM card is a basic SIM card with added PKI functionality. An RSA co-processor is added for public key-based encryption and signing with private keys. The PKI SIM card stores private keys and certified public keys needed for digital signatures and encryption (Setec, 2006).

USIM

A USIM card is a SIM used in 3G mobile telephony networks. The physical size is the same as for a GSM SIM card, but hardware is different. USIM is actually an application running on a UICC (universal integrated circuit card) storing a pre-shared secret key (Lu, 2002).

ISIM

An ISIM card consists of an application (ISIM) residing on a UICC. ISIM provides secure authentication of handheld users to IMS (IP multimedia system) services (Dietze, 2005).

Wireless Security Protocols

Security protocols are—for wired networks—implemented by (Perelson & Botha, 2004): authentication services, confidentiality services, non-repudiation services, and authorization. Four wireless security protocol types are needed:

- access control to mobile devices,
- local access control to networks and network services,
- remote access control to networks and network services, and
- protection of data communication to/from mobile devices.

Different protocols are presented in Markovski and Gusev (2003).

Access Control to Mobile Devices

Access control must be implemented on a mobile device itself to prevent unwanted access to confidential data stored in the device (see Figure 1). Authentication confirms a claimed user identity.

PIN and Password Authentication

A PIN is four digits from a 10-digit (0-9) keypad. However, PINs are susceptible to shoulder surfing or to systematic trial-and-error attacks due to their limited length and alphabet. Passwords are more secure than PINs since their length and alphabet are larger (Jansen, 2003).

Visual and Graphical Login

Visual authentication means that a user must remember image sequences to authenticate to a mobile device. A picture password system can be designed to require a sequence of pictures or objects matching a certain criteria and not exactly the same pictures. For example, the user must find a certain number of objects with four sides. This makes the shoulder surfing quite difficult (Duncan, Akhtari, & Bradford, 2004).

Biometrics

Biometric user authentication is a hardware solution for examining one or more physical attributes of an authorized user. Biometric controls, such as fingerprints, are becoming more common in handheld devices (Perelson & Botha, 2004).

Authorization

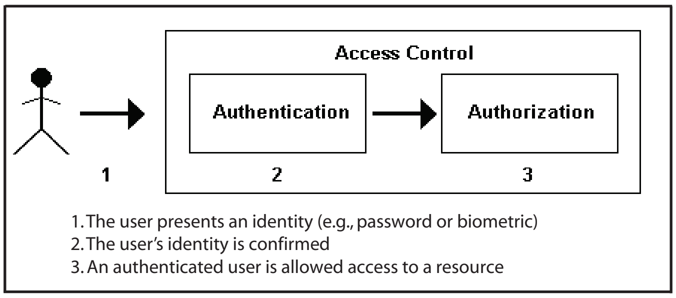
Usually mobile devices are personal, and authentication infers that the user is authorized. A corporate handheld device may however be used by several employees and may contain confidential company information. Needed user authorization features for such mobile devices include (Perelson & Botha, 2004):

- **File Masking:** Some files cannot be viewed by unauthorized users.
- **Access Control Lists:** User-related object permissions.
- **Role-Based Access Control:** User role-related permissions.

Local Network Access

Local network access protocols depend on the wireless access network type (WLAN, Bluetooth, Cellular Network, etc.). A

Figure 1. The access control principle



1. The user presents an identity (e.g., password or biometric)
2. The user's identity is confirmed
3. An authenticated user is allowed access to a resource

WLAN is usually an access network to a LAN. Authentication for LAN resources is thus also needed unless WLAN authentication is integrated in a single-sign-on scheme. Local network access protocols are described in later sections.

Remote Network Access

Secure remote network access from a mobile device requires a VPN (virtual private network), which is a protected data path in an existing unsecured network to a private LAN. VPNs can be based on different protocols: IPSec (IP security), SSL (secure socket layer), or SSH (secure shell).

IPSec VPN

IPSec operates at the network layer of the OSI model. IPSec protocols are:

- ESP (encapsulating security payload) for authentication, data confidentiality, and message integrity;
- AH (authentication header) for authentication and message integrity; and
- IKE (Internet key exchange protocol) for encryption key exchange.

IPSec VPNs require VPN client software in mobile devices (Davis, 2001).

SSL VPN

The encrypted tunnel is established at the session layer of the OSI model. SSL VPN clients communicate with the VPN gateway using an SSL-supported application such as a Web browser or e-mail client. No separate VPN client software is therefore needed (Steinberg & Speed, 2005).

SSH

SSH (secure shell) is a protocol for login to and executing commands on a remote UNIX computer. SSH provides between two communicating hosts an encrypted communication channel, which can be used for port forwarding with VPN functionality (Barret et al., 2005).

Protection of Data Communication

Security protocols for protection of wireless data communication are integrated in protocols for local and remote access to networks/network services. In a cellular network a shared secret session key created by the authentication protocol is used for encryption/decryption of data communication. In a WLAN, the TKIP (temporal key integrity protocol) is integrated in the WPA security protocol, and AES (advanced encryption standard) is integrated in the WPA2 security protocol. The remote access protocols IPSec, SSL/TLS, and SSH also provide end-to-end protection of data communication with secure symmetric encryption algorithms and shared secret session keys created during authentication.

PLATFORMS FOR INTEGRATED ARCHITECTURES

Software signing and binary trust-models do not provide adequate protection against third-party programs. Fine-grained software authorization is emerging into mobile units. Typical examples include Java sandboxing and Symbian platform security. Software-based mobile platform examples are Java Mobile Environment, Symbian OS, Embedded Linux, Windows Mobile, Brew (Binary Runtime for Wireless), Blackberry OS, and Palm OS.

OS implementation vulnerabilities still remain a challenge. Integrated solutions have been proposed for executing trusted code and for secure boot. Standardization efforts are under development (e.g., Trusted Computing Group and Trusted Mobile Platform). There are different embedded on-chip security solutions, but mostly the security solution relies on combining hardware and software. Platform security examples are Texas Instruments OMAP™ Platform (Sundaresan, 2003) and Intel Wireless Trusted Platform (Intel Corporation, 2006b).

The TI platform relies on three layers of security: application layer security, operating system layer security, and on-chip hardware security. The main security features are:

- A *secure environment* provides secure execution of critical code and data by *secure mode, secure keys, secure ROM,* and *secure RAM.*
- *Secure boot/flash* prevents security attacks during device flashing/booting.
- *Run-time security* is included for security-critical tasks like encryption/decryption, authentication, and secure data management.
- A *hardware crypto engine* is also included for DES/3DES, SHA1/MD5, and RNG with two configuration modes: secure mode and user mode.

Intel platform building blocks are performance primitives (hardware) and cryptographic primitives (optimized software) for security services. Platform components include

- *trusted boot ROM* integrity validation and booting to a correct configuration;
- *wireless trusted module* processing secrets;
- *security software stack* enabling access to platform resources through standard cryptographic APIs;
- *protected storage* in system flash for secrets; and

- *physical protection* by security hardware in a single device and discrete components in a single physical package.

WIRELESS APPLICATION SECURITY

The risks described above should be addressed in wireless application design. Wireless application security includes (Umar, 2004): application access control, client/server communications security, and anti-malware protection.

Application Access Control

Many mobile platforms lack support for individual user accounts and for operating system-level logon. Mobile applications handling confidential data should require user authentication before application access is granted. In case a mobile device is lost or stolen while the device user is logged in to an application, the application should also support "session timeout." This means that a limited inactive time is specified for an application before re-authentication is required (Intel Corporation, 2006a).

Client/Server Communication Security

Typical wireless Internet connections are:

1. the wireless connection between a mobile device and an access device, and
2. the Internet connection between the mobile device and the Internet host/server via the access device.

Internet connection security should be provided at the application level.

For Web-based client/server applications, the SSL protocol provides encryption and signing of transmitted data. SSL application examples are:

- Web browsers for secure communications with Web servers,
- e-mail client software for secure reading of e-mail messages on e-mail servers, and
- SETs (secure electronic transactions) for secure financial transactions with credit cards.

For applications using customized protocols, security protocols are also customized. Alternatively, VPN techniques can be used.

Anti-Malware Protection

Most current mobile operating systems lack memory space protection. Malware can access and steal application data, such as credit card information stored in memory by wireless applications. Time and space for sensitive data in memory should be minimized (Intel Corporation, 2006a).

SECURITY OF MOBILE TECHNOLOGIES

A taxonomy of mobile technologies is:

- wireless cellular networks (GSM, DECT, GPRS, and UMTS),
- wireless long-range networks (WiMax, Satellite Communication Technology),
- wireless local area networks (WLAN, ZigBee^{TM}), and
- wireless short-range networks (Bluetooth, Wireless USB).

Wireless Cellular Networks

First Generation

First-generation cellular systems, such as AMPS (advanced mobile phone system) introduced in the early 1980s, use analog transmission and provide no security.

Second Generation

2G cellular systems, such as GSM (Global System for Mobile Communications) introduced in the late 1980s and DECT, use digital transmission.

GSM security is based on a unique IMSI (International Mobile Subscriber Identity) and a unique secret key (Ki) stored in the SIM card of each subscriber. The Ki is never transmitted over the network. Every GSM network has:

- *AUC (authentication center),* a protected database containing a copy of Ki;
- *HLR (home location register)* for subscriber information;
- *VLR (visitor location register)* for information of each mobile station currently located in the geographical area controlled by the *MSC (Mobile Station Controller)*; and
- *EIR (equipment identity register)* for lists of mobile stations on the network. Stations have unique IMEI (International Mobile Equipment Identity) numbers.

When a mobile station enters a GSM network for the first time, the IMEI is transmitted for determination in which AUC/HLR subscriber data is stored. The MSC/VLR of the visited network asks for and stores a security triplet (a unique random number RAND, a signed response SRES, a ciphering key Kc) from the AUC/HLR. SRES and Kc are calculated from RAND with Ki.

Subscriber authentication:

- RAND is sent to the mobile station.
- SRES' and Kc' are calculated from RAND with Ki.
- SRES' is sent back to MSC/VLR.
- Authenticated if SRES=SRES'.

Kc=Kc' is used for radio link encryption/decryption.

After the initial registration, IMSI is stored in the VLR. A TMSI (temporary mobile subscriber identity) is generated, transmitted back to the mobile station, stored in the SIM card, and used for future subscriber identification in the visited network.

DECT is a cellular system and a common standard for cordless telephony, messaging, and data transmission standardized by ETSI (European Telecommunications Standards Institute). DECT is similar to GSM, but cell ranges are shorter (DECT, 2006).

DECT uses several advanced digital radio techniques for efficient radio spectrum utilization. It enables high speech quality and security with low radio interference risks and low-power technology. Mobility management, responsible for DECT communication security, consists of procedures for *identity, authentication, location, access rights, key allocation, parameter retrieval,* and *ciphering* (Umar, 2004).

2.5 Generation

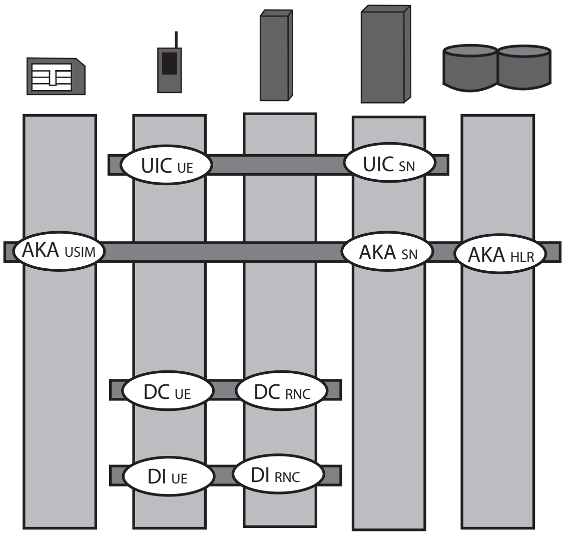
The GPRS (2.5G) infrastructure equals GSM. TSMI is replaced by P-TMSI (packet TMSI) and by P-TMSI signature as alternate identities. Mapping between IP addresses and IMSI is generated in the HLR GPRS Register. GPRS authentication is performed by SGSN (serving GPRS support node). As a consequence, user data and signaling are encrypted all the way from the mobile station to the SGSN. Tunneling, firewalls, and private IP techniques are used. IP addresses are assigned after authentication and encryption algorithm negotiations.

Third Generation

UMTS, Universal Mobile Telecommunications System, a standard for third-generation (3G) systems for mobile communication, referred to as International Mobile Telecommunications 2000 (IMT-2000) and initiated by the International Telecommunication Union (ITU), is presently being developed by the Third Generation Partnership Project (3GPP).

The UMTS security architecture is based on 2G/2.5G security. Some GSM security features have been improved and some new features have been added. The UMTS security mechanisms are (see Figure 2): user identity confidentiality (UIC), authentication and key agreement (AKA), confidentiality of user and signaling data (DC), and integrity of signaling data (DI). See Lu (2002) for UMTS security details.

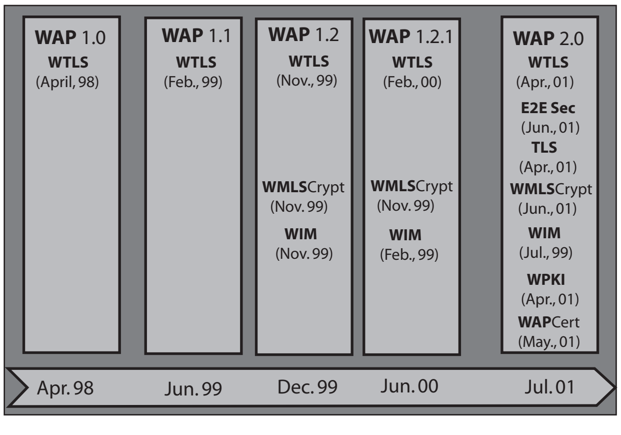
Figure 2. UMTS functional security architecture; UE is user equipment and RNC is radio network controller



WAP

WAP (wireless application protocol) is an open mobile device application standard. WAP security protocols and specifications are being developed by the WAP Forum (Open Mobile Alliance, 2006). The evolution of WAP security specifications is shown in Figure 3.

Figure 3. The development of WAP security specifications



WTLS/TLS/SSL

SSL/TLS are TCP-based security protocols for communication in client/server applications. WAP 2.0 adopts TLS as security protocol and supports the tunneling of SSL/TLS sessions through a WAP/WAP proxy. TLS/SSL in WAP 2.0 is a complement to the similar UDP-based WTLS protocol in earlier WAP versions. Server authentication and mutual authentication are options in WTLS/TLS/SSL-protected WAP applications.

WMLScript Crypto Library, WIM, and WPKI

The lack of non-repudiation services and end-user authentication was addressed in WAP 1.2. The WMLScript (Wireless Markup Language Script) Crypto Library provides cryptographic functionality for WAP clients. WAP identity module (WIM) is used in WTLS and application-level security functions. A WIM stores and processes user authentication information, such as private keys. A WIM implementation example is a mobile phone S/WIM card (combined SIM and WIM). WPKI (wireless public key infrastructure) is a mobile environment PKI supported since WAP 2.0 (Open Mobile Alliance, 2006).

i-mode

i-mode is a Japanese competitor to WAP for m-commerce. i-mode security features are:

- protection of the radio link between the i-mode handset and the base station,
- encryption/authentication of data transmitted between i-mode mobile devices and Web sites, and
- protection of private network links between the i-mode center and special service providers like banks.

The radio link is protected using SSL and other protocols, which are not public. Security of Web site connections and private network links are based on SSL. Mutual certificate authentication is supported (Umar, 2004).

Bluetooth

Bluetooth provides wireless short-distance transmission of data and voice signals between electronic devices. The specifications are defined by Bluetooth SIG (2006). The security is based on *authentication, authorization,* and *encryption.* The *security modes* are:

1. no security measures,
2. security measures based on authorization, and
3. authentication and encryption.

Table 1. Bluetooth service levels

	Authorization	Authentication	Encryption
Trusted	Yes	Yes	Yes
Untrusted	No	Yes	Yes
Unknown	No	No	Yes

Authentication

Bluetooth device authentication is a unidirectional or mutual challenge/response process. Secret keys, called *link keys,* are generated either dynamically or by pairing. For dynamic link key generation, a passkey—the same passkey—must be entered in both connecting devices each time a connection is established. In pairing, a long-term stored link key is generated from a user-entered passkey, which can be automatically used in several connection sessions between the same devices.

Authorization

In authorization, a Bluetooth device determines whether or not another device is allowed access to a particular service. Levels of trust are *trusted, untrusted*, or *unknown.* Service levels are shown in Table 1.

Encryption

Bluetooth data transmission uses 128-bit encryption. Encrypted data can only be viewed by a device owning the proper decryption key. The encryption key is based on the link key.

ZigBee

ZigBee is a low-cost, low-power communications standard for wireless data communication in home and building automation. The ZigBee stack architecture is based on the standard OSI model. The IEEE 802.15.4-2003 standard defines the physical (PHY) layer and the medium access control (MAC) sub-layer. The ZigBee Alliance builds on this foundation by providing the network (NWK) layer and a framework for the application layer with: the application support sub-layer (APS), ZigBee device objects (ZDO), and manufacturer-defined application objects.

Security services are defined for key establishment, key transport, frame protection, and device management. The MAC, NWK, and APS layers are responsible for the secure transport of their respective frames. Data encryption uses the symmetric key 128-bit AES algorithm. Frame integrity is protected, since frames cannot be modified by parties without cryptographic keys. Replayed data frames are rejected by a frame freshness verification function of the NWK layer. Furthermore, the APS sub-layer establishes and maintains security relationships. ZDO manages the security policies and the security configuration of a device. Access control uses a list of trusted devices maintained by a ZDO (ZigBee Alliance, 2004).

WLAN

Broadband mobile communication is supported by a WLAN, which gives mobile users LAN connectivity through a high-speed radio link. Major WLAN security standards are (Pulkkis, Grahn, Karlsson, Martikainen, & Daniel, 2005): IEEE 802.11/WEP, WPA, and IEEE 802.11i.

WEP is not recommended due to security flaws. Data encryption is based on static encryption keys, and no user authentication mechanisms are specified. *WPA* addresses the WEP vulnerabilities and is based on IEEE 802.11i (see Figure 4).

The main features of WPA are:

- Temporal key integrity protocol (TKIP) to provide dynamical and automatically changed encryption keys, and
- IEEE 802.1X and EAP (extended authentication protocol) to provide strong user authentication.

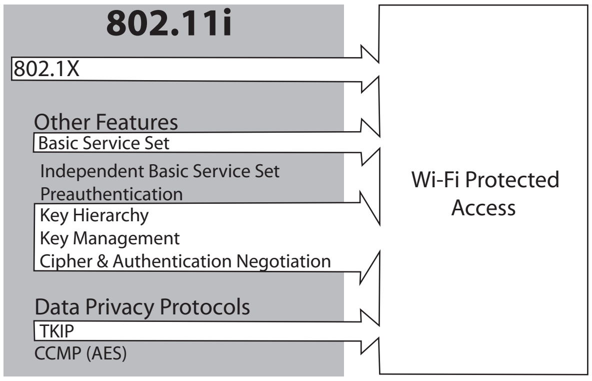
CCMP (cipher block chaining message authentication protocol) is an IEEE 802.11i protocol that uses the AES (advanced encryption standard) to provide stronger encryption than TKIP.

WiMax

WiMax is a new technology for wireless broadband Internet access. The MAC layer of the WiMax network stack has a security sub-layer with (Puthenkulam & Yin, 2005):

- a base station device and mobile user authentication capability based on the EAP protocol, X.509 certificates, and AAA servers (Radius, Diameter);
- encryption key management using the privacy key management protocol (PKM) v2;
- AES-CCM authenticated encryption of all data communication—the Encryption Key Refresh Mechanism supports high data rates; and
- CMAC (cipher-based message authentication code) and HMAC (hash-based message authentication code), which handle control message integrity protection.

Figure 4. IEEE 802.11i features in WPA



Wireless USB

An USB wire provides two security services: (1) a wanted interconnection of two devices is created, and (2) all data in transit is protected from casual observation or malicious modification by external parties.

The goal of Wireless USB security is to provide analogous security services. Hosts and wirelessly connected devices are required to authenticate each other to avoid man-in-the-middle attacks. Data communication between a host and a wirelessly connected device is confidential and integrity-checked by AES-128/CCM encryption. Secret encryption keys are shared by mutually authenticated hosts and wirelessly connected devices (Wireless, 2005).

Satellite Communication Technology

A communication satellite permits two or more earth stations to send radio messages to each other over far distances. For satellite communication security it is necessary that earth stations have significant physical security, and RF (radio frequency) communication channels between satellites and earth stations are protected.

Satellite communications are normally secured by scrambling satellite signals using cryptography or transmitting same signals over several frequencies. The data bits are basically transmitted on different signals based on a secret scheme. The receiver of a signal must thus be aware of the secret scheme. Additional security protocols like IPSec can be used to encrypt radio messages. However, such protocols slow down data transmission. The main challenge is thus to find a good balance between performance and security (Umar, 2004).

FUTURE TRENDS

Privacy, security, and trust issues are and will be of major importance. The growth of the Internet and m-commerce will dramatically increase the amount of personal and corporate information that can be captured or modified. In the near future ubiquitous computing systems will accentuate this trend. We can likewise expect an increase in privacy and security risks, not only with the emergence of mobile and wireless devices, but also with sensor-based systems, wireless networking, and embedded devices. Ubiquitous computing technologies will probably suffer from the same sorts of unforeseen vulnerabilities that met the Internet society.

CONCLUSION

Mobile terminals face security threats due to openness. Platforms are open for external software and content. Malicious software, like Trojan horses, viruses, and worms, has started to emerge. Fine-grained software authorization has been proposed. Downloaded software may then access particular resources only through user authorization. OS implementation vulnerability still remains a challenge because of difficulties in minimizing OS code running in privileged mode. Integrated hardware solutions may be the solution.

Wireless security architectures have many options, and many standards/protocols addressing wireless security are quite recent, especially standards/protocols based on public key cryptography. Therefore more practical experience from the use of these protocols/standards in mobile computing is needed for reliable estimation of the provided security.

REFERENCES

Barrett, J.D., Silvermann, E.R., & Byrnes, G.R. (2005). *SSH, the secure shell: The definitive guide* (2nd ed.). O'Reilly.

Barun, T., & Danzeisen, M. (2001). Secure mobile IP communication. *Proceedings of the IEEE 26th Annual Conference on Local Computer Networks* (pp. 586-593).

Bluetooth SIG. (2006). *The official Bluetooth wireless info site.* Retrieved August 8, 2006, from http://www.bluetooth.com

Calhoun, P., Johansson, T., Perkins, C., Hiller, T., & McCann, P. (2005, August). *Diameter mobile IPv4 application.* IETF, RFC 4004.

Davis, C. (2001). *IPSec: Securing VPNs.* New York: McGraw-Hill.

DECT Forum. (2006). Retrieved August 8, 2006, from http://www.dect.org

Dietze, C. (2005). The smart card in mobile communication: Enabler of next-generation (NG) services. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities.* Hershey, PA: IRM Press.

Duncan, M. V., Akhtari, M. S., & Bradford, P. G. (2004). Visual security for wireless handheld devices. *JOSHUA—Journal of Science & Health at the University of Alabama, 2.*

Handheld Security. (2006). *Laura Taylor, part I-V (2004-2005).* Retrieved August 8, 2006, from http://www.firewall-guide.com/pda.htm

Hwu, J.-S., Chen, R.-J., & Lin, Y.-B. (2006). An efficient identity-based cryptosystem for end-to-end mobile security. *IEEE Transactions on Wireless Communication.*

Intel Corporation. (2006a). *Wireless application security: What's up with that?* Retrieved August 8, 2006, from http://www.intel.com/cd/ids/developer/asmo-na/eng/57399.htm?page=1

Intel Corporation. (2006b). *Intel wireless trusted platform: Security for mobile devices.* Retrieved August 8, 2006, from http://www.intel.com/design/pca/applicationsprocessors/whitepapers/300868.htm

ISO/IEC 7498-1. (1994). *Information technology—Open systems interconnection—Basic reference model: The basic model, 1994.*

ISO 7498-2. (1989). *Information processing systems—Open systems interconnection—Basic references model—Part 2: Security architecture, 1989.*

Jansen, W. A. (2003, May 12-15). Authenticating users on handheld devices. *Proceedings of the 15th Annual Canadian Information Technology Security Symposium (CITSS)*, Ottawa, Canada. Retrieved August 8, 2006, from http://csrc.nist.gov/mobilesecurity/publications.html#MD

Lu, W.W. (2002). *Broadband wireless mobile, 3G and beyond.* New York: John Wiley & Sons.

Markovski, J., & Gusev, M. (2003, April). Application level security of mobile communications. *Proceedings of the 1st International Conference Mathematics and Informatics for Industry (MII 2003)* (pp. 309-317), Thessaloniki, Greece.

Olzak, T. (2005). *Wireless handheld device security.* Retrieved August 8, 2006, from http://www.securitydocs.com/pdf/3188.PDF

Open Mobile Alliance. (2006). *WAP forum.* Retrieved August 8, 2006, from http://www.wapforum.org/

Perelson, S., & Botha, R. (2004, July). An investigation into access control for mobile devices. In H. S. Venter, J. H. P. Eloff, L. Labuschagne, & M. M. Eloff (Eds.), *Proceedings of the ISSA 2004 Enabling Tomorrow Conference on Information Security,* South Africa.

Perkins, C., & Calhoun, P. (2000). *Mobile IPv4 challenge/response extensions.* IETF, RFC 3012.

Pulkkis, G., Grahn, K., Karlsson, J., Martikainen, M., & Daniel, D. E. (2005). Recent developments in WLAN security. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities.* Hershey, PA: IRM Press.

Puthenkulam, J., & Yin, H. (2005). *802.16e: A mobile broadband wireless standard.* Broadband Wireless Division, Mobility Group, Intel Corporation. Retrieved August 8, 2006, from http://www.ewh.ieee.org/r6/scv/comsoc/0512.zip

Rankl, W., & Effing, W. (2003). *Smart card handbook* (3rd ed.). New York: John Wiley & Sons.

Setec Portal. (2006). Retrieved August 8, 2006, from http://www.setec.fi

Steinberg, J., & Speed, T. (2005). *SSL VPN: Understanding, evaluating and planning secure, Web-based remote access.* Birmingham, UK: Packt Publishing.

Sundaresan, H. (2003). *OMAP™ platform security features.* Retrieved August 8, 2006, from http://focus.ti.com/pdfs/wtbu/omapplatformsecuritywp.pdf

Umar, A. (2004). *Mobile computing and wireless communications.* Middlesex, NJ: Nge Solutions.

Wireless Universal Serial Bus Specification. (2005, May 12). *Revision 1.0.* Retrieved August 8, 2006, from http://www.usb.org/developers/wusb/docs/WUSBSpec_r10.pdf

Zao, J., Kent, S., Gahm, J., Troxel, G., Condell, M., Helinek, P., Yuan, N., & Castineyra, I. (1999). A public-key based secure Mobile IP. *Wireless Networks, 5*(5), 393-390.

ZigBee Alliance. (2004, December 14). *ZigBee™ Specification v1.0.* Retrieved August 8, 2006, from http://www.zigbee.org

KEY TERMS

Bluetooth: A technology standard for wireless short distance communication.

DECT: A cellular system and a common standard for cordless telephony, messaging. and data transmission standardized by ETSI (European Telecommunications Standards Institute).

Mobile IP: Mobile Internet protocol for IP number preservation of a mobile computer.

USIM: A SIM used in 3G mobile telephone networks.

WiMax: A technology standard for wireless broadband Internet access.

ZigBee™: A low-cost, low-power communication standard for wireless data communication in home and building automation.

Semantic Caching in a Mobile Environment

Say Ying Lim
Monash University, Australia

INTRODUCTION

Mobile computing environments enable the database servers to disseminate data via wireless channels to multiple mobile clients (Chung & Kim, 2001). It has increased popularity with the emerging trend of wireless network and usage of handheld devices, such as PDAs and other portable electronic devices. The typical nature of a mobile environment would include low bandwidth and low reliability of wireless channels, which causes frequent disconnection to the mobile users. Hence due to the constraints of the nature of mobile environment it is important to enhance the performance of the query processing, as well as improve the availability of querying particular data items especially during disconnection (Imelinkski & Korth, 1996; Malladi & Davis, 2002). Often, mobile devices are associated with low memory storage and low power computation and with a limited power supply (Myers & Beigl, 2003). Hence, it is important to help mobile clients to save the usage of its battery.

By introducing data caching into the mobile environment, it is believed to help improve data availability in case of disconnection by retrieving data that has been previously cached in the local memory and to be able to save power by having a lower data transmission. Generally, in data caching, it means the data is cached in the memory storage of the mobile device, and whenever the mobile users want to issue a query, it will first search its cache and if there exists a valid copy in the cache it returns the results immediately. Otherwise the mobile users would attempt to obtain the data from the server either using the server or broadcast strategy. Caching has emerged as a fundamental technique, especially in distributed systems, as it not only helps reduce communication costs but also off loads shared database servers.

In this article, we describe the use of caching, which allows coping with the characteristic of the mobile environment. We concentrate particularly on semantic caching, which is basically a type of caching strategy that is content-based reasoning ability with the ability to—in addition of caching query results—remember the queries that generated these results. Semantic caching provides accurate, semantic description of the content of the cache.

BACKGROUND

The effect of having the ability to cache data is of great importance, especially in the mobile computing environment than in other computing environments. This is due to the reason that contacting the remote servers for data is expensive in the wireless environment and, with the vulnerability to frequent disconnection, can further increase the communication costs (Leong & Si, 1997). There are many different types of caching strategies that serve the purpose to improve query response time and to reduce contention on narrow bandwidth (Zheng, Lee, & Lee, 2004). Caching mechanisms need to retain the frequently accessed data locally in the mobile device storage to be able to allow users to access server database queries at least partially in cases of disconnections. Hence, the more effective the caching mechanism is in keeping the frequently accessed data will result in more queries that can be served during disconnection.

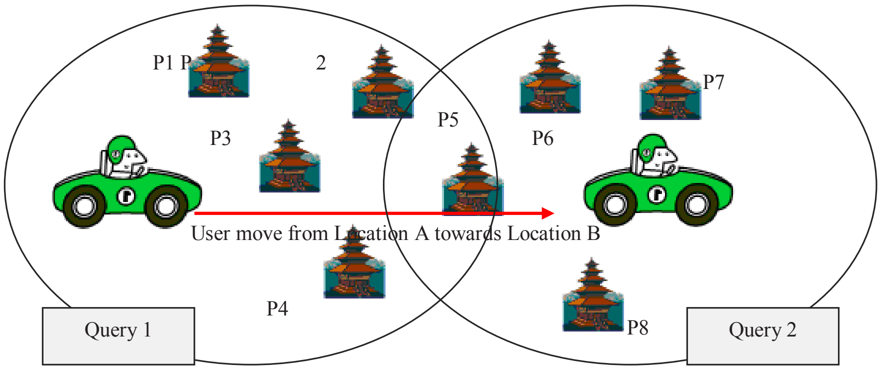
Due to limitations such as cache space, cache replacement and cache granularity, as well as cache coherence, are the three main issues that characterize caching mechanism. In traditional cache replacement, the most important factor affecting cache performance is the access probability. This refers to replacing the data with the least access probability to free up more cache space for the new data. There is a large variety of caching replacement policies and most of them utilize access probability as the primary factor in determining which data items are to be replaced.

Cache granularity relates to determining a physical form of cached data items. It appears to be one of the key issues in caching management systems. There are three different levels of caching granularities in object-oriented databases, which includes: (a) attribute caching, (b) object caching and (c) hybrid caching (Chan, Si, & Leong, 1998). Attribute caching refers to frequently accessed attributes that are stored in the client's local storage. As for object caching, instead of the attribute itself being cache, the object is cached. In attribute caching, it creates undesirable overheads due to the large number of independent cache attributes. Thus, hybrid caching, which appears to be a better approach, comprises of the combinations of both granularities.

Cache coherence—or known as invalidation strategy—involves cache invalidation and update schemes to invalidate

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Overlapping results from two queries issued



and update out-dated or non valid cached items (Chan, Si, & Leong, 1998; Cao, 2003). After a certain period, a cached data may appear as no longer valid and therefore mobile users should obtain a newer cache before retrieving the data (Xu, Tang, & Lee, 2003). There are several techniques that have been proposed to overcome this issue. These include (a) stateful server, (b) stateless server, (Barbara & Imielinski, 1994) and (c) leases file caching mechanism (Lee, Leong, & Si, 2001). Stateful server refers to the server having an obligation to its clients, which means the server has the responsibility in notifying the users about changes, if there are any. In contrast, stateless server refers to the server not aware for its clients, whereby the server broadcasts a report that contains the updated item either asynchronously or synchronously. The leases files mechanism, which is also known as lazy invalidation approach, assigns each mobile user to be responsible for invalidating its cached items.

Consequently, a good caching management strategy is needed to deal with the critical caching issues, such as caching replacement, caching granularity and caching coherence.

SEMANTIC CACHING

A better way of query processing specifically for use in a mobile environment is by allowing the users to specify precisely what data items are missing from its local storage to server the query. This could be achieved by having the previously evaluated query results being cached (Dar et al., 1996; Roussopoulos, 1991).

Using Semantic Caching in a Mobile Environment

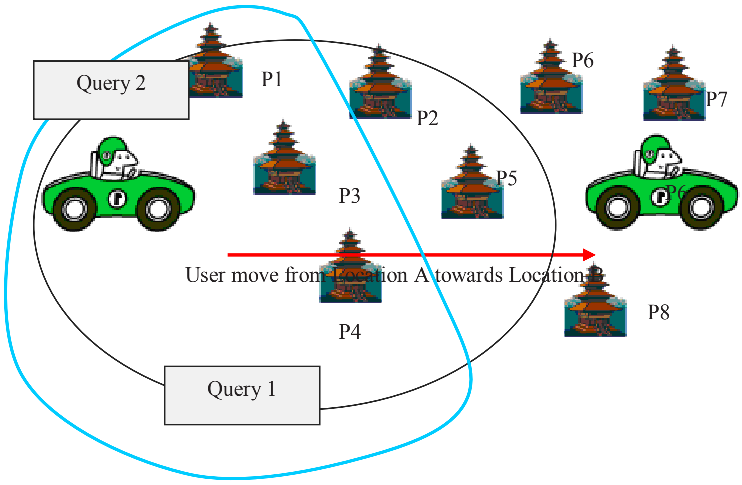
A semantic cache is defined as consisting of a set of distinct semantic segments, which can be decomposed into separate components or come together as a whole of the query results. A semantic segment *S* can be specified by having <*SR*,*SA*,*SP*,*SC*> whereby *SR* and *SA* define the base relation of relation and attributes in the creation of the semantic segment respectively. *SP* is to indicate the criteria that *S* satisfies, and *SC* indicates the actual content of *S*, which is represented by pages. (Ren, Dunham, & Kumar, 2003)

Semantic caching stores semantic descriptions and associated answers of the previous queries in the mobile client (Dar et al., 1996). The main feature of semantic caching is the content-based reasoning ability as well as the fact that only the required data, as opposed to a file or pages of data, is transmitted over the wireless channel.

When a new query exists, the mobile client can determine whether should it be totally answered by how much can it be answered and what data are missing. With these abilities, the wireless traffic can be greatly reduced because only the needed data are transferred. This helps with disconnection too, since total or partial results may be obtained even when the server is unreachable (Lee, Heong, & Si, 1999). As a result, if a query can be partially answered from the cache, the volume of missing data requested from the server as well as the wireless bandwidth consumed can be reduced. And if the query could be answered completely based on the cache, then no communication between the client and the server is required at all. This ability is of particular significance during disconnection, which is the main constraint the mobile environment is currently facing. This also leads to reduction of overhead due to redundant computation as the amount of data transferred over the wireless channel can be substantially reduced.

Example 1: Suppose a mobile user who is traveling from one location to another location suddenly wished to find a nearby rest place. So the user issue a query while he is in Location *A* and the server returns the nearest rest place which is *P*1, *P*2, *P*3, *P*4, *P*5. But the user is not satisfied with the results. So he re-issued another query while he is moving

Figure 2. Issuing a 3NN query



towards Location *B*. And the query returns another set of results which may contain some overlapping results such as the same *P*5 as the user previously received. Hence, the cached results are immediately returned since it has been previously cached. Thus, the user actually only needs to submit the complement of the new query in order to obtain only results that are not the same as the one previously obtained. This example can be illustrated as in Figure 1. This shows that semantic caching not only saves the wireless bandwidth due to less retransmission, but also reduces the query response time since some cached results can be immediately returned.

Benefits and Limitations

There are several advantages that can be gained by using semantic caching, with the main reason that because only required data are being transferred communication cost between the client and the servers would be reduced. Moreover, cache space overhead is low for semantic caching since only the data that satisfy previous queries are being stored. With the ability of semantic caching to keep semantic information, it enables missing data to be exactly determined, which causes easy parallel query processing. Hence, semantic caching is very efficient to be used in the mobile environment since more autonomy is given to the clients and partial results can be derived when disconnections from the wireless channels occur (Ren, Dunham, & Kumar, 2003).

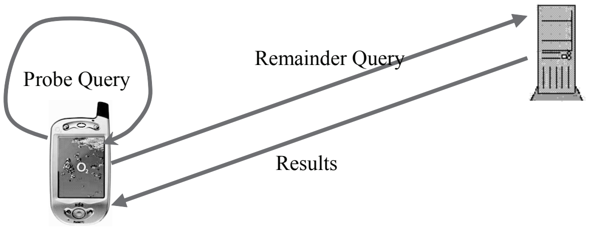
Besides all the benefits semantic caching brings in, there are also limitations and drawbacks that semantic caching brings. Generally, semantic caching captures the semantics of the queries only, and ignores the semantics of the cached objects. Therefore, the granularity is at a query level that helps answering similar queries faster, but cached objects from different types of queries become difficult. (Hu et al., 2005) In addition, the types of spatial queries supported by semantic caching are rather limited to simple range query and nearest neighbor (NN) query (Ren & Dunham, 2003; Zheng & Lee, 2001). It is difficult to support complex queries such as *k*-nearest-neighbor (*k*NN). Besides that, it also demonstrates complicated cache management. For example, when a new query to be cached overlaps some cached query, a decision has to be made whether to bring these two queries or to trim either of them. When the cache size grows, all these drawbacks would become more remarkable.

Example 2: The same scenario as described in Example 1 but, instead of the user issuing a query 2 in a new location after the query 1 which has been sent, he would like to issue query 2 that is comprised of 3 nearest neighbor (*3*NN) query. Due to the limitation of semantic caching that is not able to trim a *3*NN query from the first query, the user would have to sent a full complete query 2 to the server even though the results data *P*1, *P*3, *P*4 have been cached as a result of the first query that has been issued earlier on. They are actually partial results of the new query and should have been returned immediately to the user but are not able to do so. This example can be illustrated as in Figure *2*. The retransmission of this result to the mobile users has wasted the wireless bandwidth as well as unnecessary transfer and has prolonged the response time (Hu et al., 2005).

Semantic Caching in Query Processing

In order to process a query from a semantic cache, first of all we would check whether the query can be answered locally by the cache. If the answer can be obtained from the cache, the results are locally processed from the cache in the mobile device. However, in cases where the answers are not fully obtainable from the cache, but can only be partially

Figure 3. Semantic caching query processing mechanisms



answered, we will trim the original query by either removing or annotating those parts that are answered and send it over the wireless channel to the server for further processing (Godfrey & Gryz, 1999). In other words, if the answer can be totally answered from the cache a probe query is being issued, whereas a remainder query is being issued to the server when only partial answers are obtainable. Figure 3 shows the semantic caching query processing mechanisms (Waluyo, Srinivasan, & Taniar, 2005).

Example 3: Consider a mobile user who previously issued queries in getting movie information. This information had been cached into his local memory storage previously. Now he would like to send another query obtaining another list of movie information. So, it first processes the query locally via the semantic cache and determines if any of the semantic segments contribute to the query results from the cache index. If the result can be partially answered by the semantic segment then a probe query will retrieve the results that can be obtained locally, and a remainder query will be issued to the server for evaluation to define the other partial results. And then the result for the query is obtained by integrating all the partial results into a single result, which may consist of the results from probe query and remainder query.

Due to the fact that mobile users in a typical mobile environment move around frequently by changing location has opened up a new challenge of answering queries that is dependent on the current geographical coordinates of the users (Barbara, 1999). This is known as location dependent queries (Park, Song, & Hwang, 2005). An example of a location dependent query can be, "Find the nearest restaurants from where I am standing now." This is an example of static object whereby restaurants are not moving. An example of a dynamic object would be, "What is the nearest taxi that will pass by me?"

Queries should be processed in a way that minimizes the consumption of bandwidth and battery power. The problem is challenging because the user location is changing and the results would also change accordingly (Shi, Li, & Wang, 2002). And with the keep on changing location that causes changing results to be downloaded would cause high communication costs if excessive communication is needed to and from the server several times (Seydim, Dunham, & Kumar, 2001). Hence, caching plays a role in location dependent query processing. This allows the queries to be answered without connecting to the server.

In summary, there are a series of steps that can be carried out in answering a query. First of all, when a query has been issued, the local cache is checked to see whether the results can be obtainable locally or not. If there are no suitable answers to the queries that are issued that correspond to the location of the user in cases of location dependent queries, then the information of the location of the user will be transmitted to the server to answer the query and returned back to the user. Otherwise, if there is some related data from the cache itself, then the data can be retrieved directly from the cache and the answer to the query has been completed. For a location dependent situation, by giving the current location of the user as well as the speed, the time when this user will move to another location can be computed and determined. However, before the whole answering process ends, after getting the results from the server, a new cache is inserted into the local cache memory for future use (Ren & Dunham, 2000).

FUTURE TRENDS

There have been several researches done in the area of semantic caching in a mobile environment. The usage of semantic caching has obviously provoked extensive complicated issues. There are still many limitations of the nature of the mobile environment that generate a lot of attention from research in finding a good cache strategy that is specifically designed for use only in the mobile computing environment.

In the future, it is critical to design algorithms for using semantic caching to cope with the low bandwidth of the wireless channel as well as the vulnerable disconnection problem. Applying semantic caching to several different scenarios of location dependent queries, including in a multiple cell environment and a cooperative strategy between multiple clients, is also beneficial. Caching management strategies, which focus on semantic mechanisms that are designed for real mobile queries that will utilize space more efficiently, are also needed in the future.

Besides these, further investigation on other cache replacement policies as well as granularities issues, which exploit the semantics of the caching data in terms of size or access pattern, is desirable. Employing the semantic caching into a broadcasting environment, which reduces the size of a broadcast cycle and improves the tuning time, is necessary.

CONCLUSION

Although there are significant increases in the popularity of mobile computing, there are still several limitations that are inherent, be it the mobile device itself or the environment itself. These include limited battery power, storage, communication costs, and bandwidth problem. All these have become present challenges for researchers to address.

In this article we have described issues of caching in a mobile environment with its advantages and disadvantages focusing mainly on semantic caching. We include adapting semantic caching in both location and non-location dependent queries. This article serves as a valuable starting point for those who wish to gain some introductory knowledge about the usefulness of caching, particularly semantic caching.

REFERENCES

Barbara, D., & Imielinski, T. (1994, November). Sleepers and workaholics: Caching strategies in mobile environments. *MOBIDATA: An Interactive Journal of Mobile Computing*, *1*(1).

Chan, B. Y., Si, A., & Leong, H. V. (1998). Cache management for mobile databases: Design and evaluation. In *Proceedings of the International Conference on Data Engineering (ICDE)* (pp. 54-63).

Chung, Y. D., & Kim, M. H. (2001). Effective data placement for wireless broadcast. *Distributed and Parallel Databases*, *9*(2), 133-150.

Dar, S., Franklin, M., Jonsson, B., Srivastava, D., & Tab M. (1996). Semantic data caching and replacement. In *Proceedings of 22nd International Conference on Very Large Data Bases* (pp. 330-341).

Deshpande, P. M., & Ramasamy, K. (1998). Caching multidimensional queries using chunks. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp. 259-270).

Dong Jung, Y., You, H., Lee, J. W., & Kim, K. (2002). Broadcasting and caching policies for location dependent queries in urban areas. In *Proceedings of the 2nd International Workshop on Mobile Commerce* (pp. 54-60).

Franklin, M., Carey, M., & Livny, M. (1992). Global memory management on client-server architecture. In *Proceedings on International Conference on Very Large Databases* (pp. 596-609).

Ganguly, S., & Alonso, R. (1993). Query optimization in mobile environments. In *Proceedings of Fifth Workshop on Foundation of Models and Languages for Data and Objects* (pp. 1-17).

Godfrey, P., & Gryz, J. (1999). Answering queries by semantic caches. In *Proceedings of Database and Expert Systems Applications (DEXA)* (pp. 485-498).

Häkkilä, J., & Mäntyjärvi, J. (2005). Combining location-aware mobile phone applications and multimedia messaging. *Journal of Mobile Multimedia*, *1*(1), 18-32.

Hurson, A.R., & Jiao, Y. (2005). Data broadcasting in mobile environment. In D. Katsaros, A. Nanopoulos, & Y. Manolopaulos (Eds.), *Wireless information highways* (Chapter 4). Hershey, PA: IRM Press.

Imielinski, T., & Badrinath, B. (1994). Mobile wireless computing: Challenges in data management. *Communications of the ACM, 37*(10), 18-28.

Imielinski, T., Viswanathan, S., & Badrinath, B. R. (1994). Energy efficient indexing on air. In *Proceedings of the ACM Sigmod Conference* (pp. 25-36).

Imielinski, T., Viswanathan, S., & Badrinath, B. R. (1997). Data on air: Organisation and access. *IEEE Transactions on Knowledge and Data Engineering, 9*(3), 353-371.

Jayaputera, J., & Taniar, D. (2005). Data retrieval for location-dependent queries in a multi-cell wireless environment. *Mobile Information Systems*, *1*(2), 91-108.

Keller, A. M., & Basu J. (1996). A predicate-based caching scheme for client-server database architectures. *The VLDB Journal, 5*(2), 35-47.

Kottkamp, H.-E., & Zukunft, O. (1998). Location-aware query processing in mobile database systems. In *Proceedings of ACM Symposium on Applied Computing* (pp. 416-423).

Lee, K. C. K., Leong, H. V., & Si, A. (2000). A semantic broadcast scheme for a mobile environment based on dynamic chunking. In *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)* (pp. 522-529).

Lee, C. K. K., Leong, H. V., & Si, A. (2001). Adaptive semantic data broadcast in a mobile environment. In *Proceedings of the 2001 ACM Symposium on Applied Computing* (pp. 393-400).

Lee, K. C. K., Leong, H. V., & Si, A. (2002). Semantic data access in an asymmetric mobile environment. In *Proceedings of the 3rd Mobile Data Management* (pp. 94-101).

Lee, K. C. K., Leong, H. V., & Si, A. (1999, April). Semantic query caching in a mobile environment. *ACM Mobile Computing and Communications Review, 3*(2), 28-36.

Lee, D. L., Hu, Q., & Lee, W. C. (1998). Indexing techniques for data broadcast on wireless channels. In *Proceedings of the 5th Foundations of Data Organization* (pp. 175-182).

Lee, W. C., & Lee, D. L. (1996). Using signature techniques for information filtering in wireless and mobile environments. *Journal on Distributed and Parallel Databases*, *4*(3), 205-227.

Lee, G., Lo, S-C., & Chen, A. L. P. (2002). Data allocation on wireless broadcast channels for efficient query processing. *IEEE Transactions on Computers*, *51*(10), 1237-1252.

Leong, H. V., & Si, A. (1997). Database caching over the air-storage. *The Computer Journal*, *40*(7), 401-415.

Lee, D-L., Zhu, M., & Hu, H. (2005). When location-based services meet databases. *Mobile Information Systems*, *1*(2), 81-90.

Lee, D. K., Xu, J., Zheng, B., & Lee, W-C. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing*, *2*(3), 65-72.

Liberatore, V. (2002). Multicast scheduling for list requests. In *Proceedings of IEEE INFOCOM Conference* (pp. 1129-1137).

Malladi, R., & Davis, K. C. (2002). Applying multiple query optimization in mobile databases. In *Proceedings of the 36th Hawaii International Conference on System Sciences* (pp. 294 -303).

Myers, B. A., & Beigl, M. (2003). Handheld computing. *IEEE Computer Magazine*, *36*(9), 27-29.

Park, K., Song, M., & Hwang C-S. (2004). An efficient data dissemination scheme for location dependent information services. In *Proceedings of the First International Conference on Distributed Computing and Internet Technology (ICDCIT 2004)* (Vol. 3347, pp. 96-105). Springer-Verlag.

Ren, Q., Dunham, M. H., & Kumar, V. (2003). Semantic caching and query processing. *IEEE Transactions on Knowledge and Data Engineering*, *15*(1), 192-210.

Ren, Q., & Dunham, M. H. (2000). Using semantic caching to manage location dependent data in mobile computing. In *Proceedings of the 6th International Conference on Mobile Computing and Networking* (pp. 210-221).

Triantafillou, P., Harpantidou, R., & Paterakis, M. (2001). High performance data broadcasting: A comprehensive systems "Perspective." In *Proceedings of the 2nd International Conference on Mobile Data Management (MDM 2001)* (pp. 79-90).

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Research on location-dependent queries in mobile databases. *International Journal on Computer Systems: Science and Engineering*, *20*(3), 77-93.

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Indexing schemes for multi-channel data broadcasting in mobile databases. *International Journal of Wireless and Mobile Computing*, *1*(6).

Xu, J., Hu, Q., Lee, D. L., & Lee W.-C. (2000). SAIU: An efficient cache replacement policy for wireless on-demand broadcasts. In *Proceedings of the 9th International Conference on Information and Knowledge Management* (pp. 46-53).

Xu, J., Tang, X., & Lee D. L. (2003). Performance analysis of location-dependent cache invalidation schemes for mobile environments. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *15*(2), 474-488.

Xu, J., Zheng, B., Lee, W-C., & Lee, D. L. (2003). Energy efficient index for querying location- dependent data in mobile broadcast environments. In *Proceedings of the 19th IEEE International Conference on Data Engineering (ICDE '03)* (pp. 239-250).

Xu, J., Hu, Q., Lee, W.-C., & Lee, D. L. (2004). Performance evaluation of an optimal cache replacement policy for wireless data dissemination. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, *16*(1), 125-139.

Yajima, E., Hara, T., Tsukamoto, M., & Nishio, S. (2001). Scheduling and caching strategies for correlated data in push-based information systems. *ACM SIGAPP Applied Computing Review*, *9*(1), 22-28.

Zheng, B., Xu, J., & Lee, D. L. (2002, October). Cache invalidation and replacement strategies for location-dependent data in mobile environments. *IEEE Transactions on Computers*, *51*(10), 1141-1153.

KEY TERMS

Caching: Techniques of temporarily storing frequently accessed data designed to reduce network transfers and therefore increase speed of download

Cache Management Strategy: A strategy that relates to how client manipulates and maintains the data that has been cached in an efficient and effective way.

Location-Dependent Queries: A type of query whose results depend on the location of the issuer whereby when the client moves around the results may change accordingly.

Mobile Computing: Mobile computing implies wireless transmission, which enables users to use a computing device while in transit anywhere, anytime.

Mobile Query Processing: Database query is being sent by mobile users through wireless communication and being processed by a station server.

Semantic Caching: A type of caching technique that has content-based reasoning ability.

Semantic Enrichment of Location-Based Services

Vassileios Tsetsos
University of Athens, Greece

Christos Anagnostopoulos
University of Athens, Greece

Stathes Hadjiefthymiades
University of Athens, Greece

INTRODUCTION

Location-based services (LBS) are considered the most popular mobile telecommunication services besides the traditional ones, for example, SMS and MMS. They are believed to constitute the *killer applications* for next generation mobile networks, since they enable adaptive location-driven content provision. Such services can be provided wherever the location of mobile users can be determined. Nowadays, there is a wide range of methods for estimating the location of users in both *indoor* (i.e., in-building areas) and *outdoor* environments (Schiller & Voisard, 2004).

Outdoor LBS are more developed than their indoor counterparts due to the existence of positioning and topological information systems, GPS (global positioning system) and GIS (geographic information system) respectively. However, almost all known LBS provide their functionality irrespectively of the actual *user context*, which may consist of user's location, physical capabilities, and/or cognitive status. Furthermore, most services ignore the semantic information of the spatial elements (e.g., stairs, elevators, and physical obstacles), other than the Euclidean distance.

In this article, we describe issues related to the development of *intelligent* and *human-centered* LBS for indoor environments. We focus on the navigation service. Navigation is probably the most challenging LBS since it involves relatively complex algorithms and many cognitive processes (e.g., combining known paths for reaching unknown destinations, minimizing path length). With the proposed system, we try to incorporate intelligence to navigation services by enriching them with the semantics of users and navigation spaces. Such semantic information is represented and reasoned using state-of-the-art semantic Web technologies (Berners-Lee, Hendler, & Lassila, 2001).

BACKGROUND

LBS offer location-aware content provision. Apparently, a key enabler of LBS is the positioning infrastructure. As far as outdoor environments are concerned, the most commonly used positioning method is GPS, which provides spatial information with high accuracy and availability at low cost. On the other hand, there exist many alternative positioning solutions for indoor spaces, but with none of them having been standardized yet. Among these solutions are: WLAN (wireless local area network) triangulation, dead-reckoning techniques (implemented with accelerometers and digital compasses), RFID (radio frequency identification) tags, and infrared/ultrasound beacons. The authors in Hightower and Borriello (2001) provide an extensive survey of indoor positioning techniques. A basic assumption for developing our system is that we have an indoor positioning system at our disposal. Such system can locate users with "adequate" accuracy.

This article deals with indoor navigation. Former indoor navigation research focused on robot navigation. As the positioning systems have matured, more effort has been put on developing indoor navigation services for pedestrians, such as museum guides aiding the sightseeing of tourists. An indicative system in this category is CyberGuide (Abowd, Atkeson, Hong, Long, Kooper, & Pinkerton, 1997). Another, more recent and more sophisticated, navigation system is Navio (Gartner, Frank, & Retscher, 2004). Navio aims at developing a route modeling ontology, which provides both outdoor and indoor routing instructions to humans by identifying and formally defining the criteria, the actions and the reference objects used by pedestrians in their reasoning for routes. However, Navio research emphasizes on location fusion (i.e., the aggregation of location information from multiple sensing elements) and user interfaces and, thus, does not contribute significantly to the issue of path selection. This latter issue is of utmost importance for human-centered LBS,

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

but it is often ignored or handled in trivial ways. In general, such systems focus on the path presentation to users and on the hardware/positioning infrastructure used.

Additionally, some systems have been developed for addressing the special needs of certain user categories, such as navigation for blind people. Such systems, however, lack a holistic approach to the navigation process. This means that their approaches are not considered general enough to address the whole range of potential application requirements. This drawback of existing solutions, as well as their deficiencies that will be identified in the following subsections, have motivated the present research in user- and space-modeling, path selection and navigation algorithms.

Navigation Algorithms

Since navigation is a path-searching algorithmic problem, the decision on the *path-searching* algorithm used is vital for the quality of the provided service. Most of the existing navigation systems, either indoor or outdoor, make use of *traditional* shortest path algorithms (e.g., Dijkstra, A-star), thus, recognizing the minimization of Euclidian distance as the only objective in the path selection process. However, such approach overlooks the significance of other objectives more relevant to the context of the user. Hence, significant research on that topic has identified that pedestrian navigation needs more sophisticated and human-centered path-searching algorithms. Authors in Duckham and Kulik (2003) have proposed the "simplest path algorithm." In this algorithm, the selected path is the one with the lowest possible complexity in navigation instructions. This work belongs to the category of approaches that introduce modifications of well-known graph routing algorithms like the aforementioned shortest path algorithms. A rather similar approach is discussed in Grum (2005), where the proposed navigation algorithm computes the "least risk path." The term *risk* refers to the possibility of the user getting lost.

The aforementioned algorithms, although providing more "intuitively-correct" paths than the conventional shortest paths algorithms, do not take into consideration the user semantics, as dictated by the modern design paradigm "Design for All" (European Institute for Design and Disability, 2005) (a.k.a., inclusive design). This paradigm promotes the design and implementation of services and products so that they can be used by any user, without any further adaptation. The implementation of such a paradigm, in the LBS domain, would lead to services that can be consumed (in an optimal way) by any user, regardless of her special characteristics.

Spatial Models and Ontologies

The quality of *path-searching* algorithms also depends on the spatial modeling of the navigation space. Many approaches have been proposed for spatial modeling with different data representations and expressiveness. Specifically, *geometric* models represent the navigation space using a certain coordination system and mainly support geometric queries (e.g., where is the nearest coffee machine?). On the other hand, *symbolic* models represent the navigation space through sets of symbols (i.e., names) and inter-symbol relationships capturing the topological semantics (e.g., part-of and overlaps spatial relations). Finally, *hybrid* models are combinations of the two former categories, aiming at maximizing the overall expressiveness of the spatial model. An interesting comparison of spatial models is presented in Leonhardt (1998). As far as indoor navigation is concerned, only a few researchers have proposed practical, yet expressive, models. To our opinion, the most important is presented in Hu and Lee (2004). It is a hybrid model, which represents the space as semantic hierarchies of "locations" and "exits" that also carry geometric information (e.g., coordinates).

The use of semantics-based spatial models results in what has been called *semantic location-based services*. However, we claim that actual semantic LBS should not only exploit semantically enriched spatial models (symbolic or hybrid), but also take into consideration the *navigation context* (i.e., user context and instantiation of spatial model). Hence, we propose a refinement of the term *semantic LBS*, or better, *human-centered LBS*, so that it supports the following requirements:

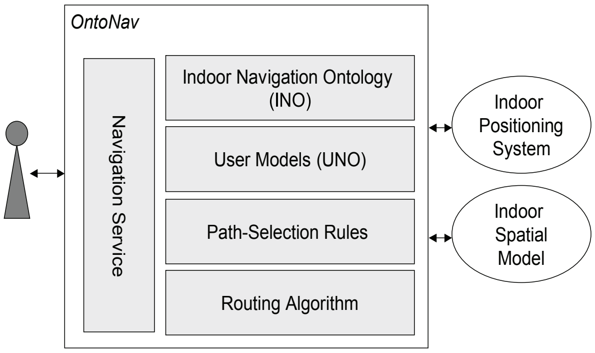
- Awareness of spatial semantics (e.g., hybrid model)
- Awareness of navigation context
- Adherence to the Design-for-All paradigm
- Reaction to dynamic user or space status changes

As will be shown in a following section, such services can be built with a knowledge-based system architecture. This architecture exploits knowledge representation methods to model the various components, and reasoning/inference techniques in order to implement the actual path selection process. The most popular and practical technology for representing models is the ontologies. Ontology is defined as "an explicit and formal specification of a shared conceptualization" (Studer, 1998, p. 185). In other words, it is a method for describing models of application domains that can be understood by machines. Knowledge reasoning is the process of inferring new implied knowledge from explicit knowledge assertions. Such reasoning can be based either on logic-based methods (i.e., resolution) or production rules (Brachman & Levesque, 2004).

SEMANTIC INDOOR NAVIGATION

In this article, we propose a framework for human-centered semantic indoor navigation, which meets the aforementioned

Figure 1. OntoNav architecture


OntoNav
Navigation Service
Indoor Navigation Ontology (INO)
User Models (UNO)
Path-Selection Rules
Routing Algorithm
Indoor Positioning System
Indoor Spatial Model

requirements. Such framework, named *OntoNav*, is based on a novel combination of practical ontology-based knowledge representation and reasoning technologies, as well as Euclidian path-searching algorithms. The architecture of the implemented navigation system is illustrated in *Figure 1* and can be decomposed to the following basic components:

Navigation Ontology: INO

This spatial ontology, named indoor navigation ontology (INO), describes the basic spatial and structural concepts of indoor environments, as well as the relationships between them. Specifically, it provides a semantic spatial model for reasoning about the selected paths. An extract of the INO taxonomy is depicted in Figure 2, illustrating a hierarchy of path elements.

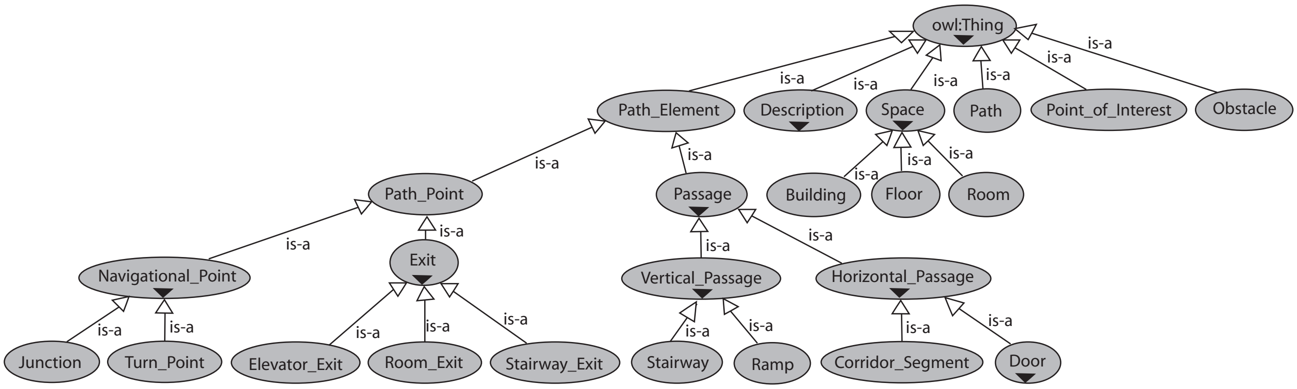
User Model: UNO

In order to describe user context (i.e., profile, capabilities, constraints and navigational preferences), we have developed a minimalistic ontology, named user navigation ontology (UNO). The concepts of such ontology represent user navigation classes. Hence, each user profile is classified into one or more navigation classes according to her characteristics. Indicative user classes are: HandicappedUser (users who cannot walk), BlindUser (users who cannot see), and LazyUser (users who always prefer elevators than stairs). Both INO and UNO have been modeled through the Web Ontology Language-OWL (McGuiness & Harmelen, 2004).

Path-Selection Rules

The path-selection process is performed through sets of production rules. The definition of such rules involves both the spatial semantics (expressed through INO) and the user semantics (expressed through UNO). The rules are applied to the INO instances in order to assert and infer which paths are considered accessible and appropriate for each user request. Moreover, the path-selection rules are further analyzed to *physical navigation rules*, *perceptual navigation rules* and *navigation preferences*. Actually, the physical navigation rules are applied first, in order to discard any paths that are not physically accessible by the user. The *perceptual navigation rules* are related to the user's cognitive status (e.g., age, education). Finally, paths that are proposed regarding the user preferences (e.g., paths containing elevators) are identified after the application of the navigation preferences. The rules are described through the Semantic Web Rule Language (SWRL) (Horrocks et al., 2004).

Figure 2. An extract of Indoor Navigation Ontology


owl:Thing
is-a
Path_Element
Description
Space
Path
Point_of_Interest
Obstacle
Path_Point
Passage
Building
Floor
Room
Navigational_Point
Exit
Vertical_Passage
Horizontal_Passage
Junction
Turn_Point
Elevator_Exit
Room_Exit
Stairway_Exit
Stairway
Ramp
Corridor_Segment
Door

Navigation Service

This service can be defined as the interface between the system and its users. It accepts navigation requests and responds with the optimal path, if any. Path optimality depends on several factors, such as suitability for current user context and length of the selected paths.

Indoor Spatial Model

The INO instances are created by a geometric representation of the indoor topology. Such geometric data may reside in a GIS as building blueprints and are transformed to actual spatial ontology instances. In our framework, it is assumed that such instances are available, (i.e., we do not deal with such data transformation issues).

Routing Algorithm

This algorithm is a central element of the framework and, in combination with the path-selection rules, is responsible for the determination of the optimal path between two given endpoints. The algorithm used in our system was a *k-shortest paths searching algorithm*. Similarly to the approach in Wu and Hartley (2004), we believe that such algorithm facilitates more flexible path selection by enabling us to incorporate additional path finding restrictions, imposed by the user profile. The main idea is that the shortest path may not be the optimal path, in general. Thus, we compute k shortest paths, since path length is always important, although it may not always be the primary selection criterion. More details on the adopted algorithm can be found in Yen (1971).

Indoor Positioning System

OntoNav symbolically locates the users in the navigation space according to the spatial model described by INO. The positioning infrastructure may vary from infrared/ultrasound beacons to WLAN triangulation or ded-reckoning techniques. It is important to note that since the positioning accuracy may be more fine-grained than the location granularity, an approximation error may be introduced in the estimated location of the user. However, this error does not significantly affect the quality of navigation.

Description of System Workflow

The end-to-end functionality of the system is depicted by means of flowcharts in the following figures. Figure 3 depicts the system initialization process, in which the spatial ontology instances are loaded. Figure 4 illustrates the workflow that takes place upon a navigation request from a user. Initially, the user registers her profile to the system, and the destination she wishes to reach. Her current location is determined through the indoor positioning system. If the user has not registered to the system again, then her profile

Figure 3. System initialization

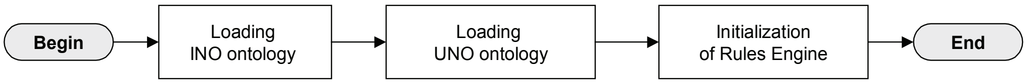

Begin
Loading INO ontology
Loading UNO ontology
Initialization of Rules Engine
End

Figure 4. System workflow after a user navigation request has been received

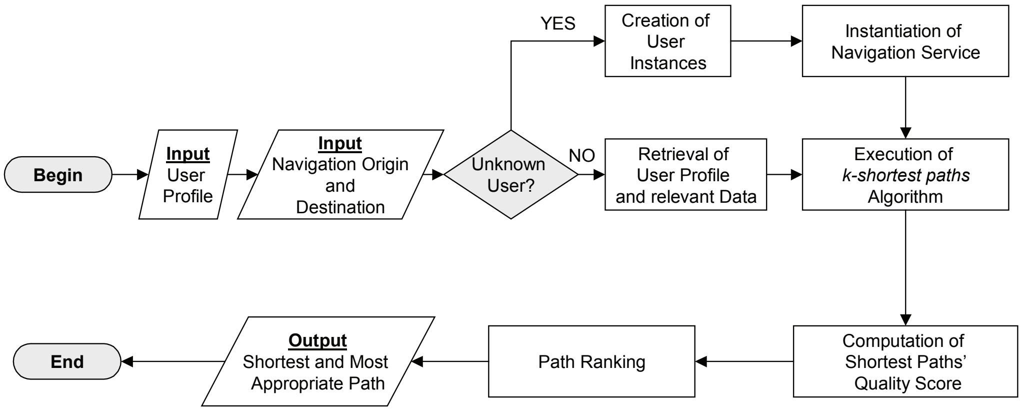
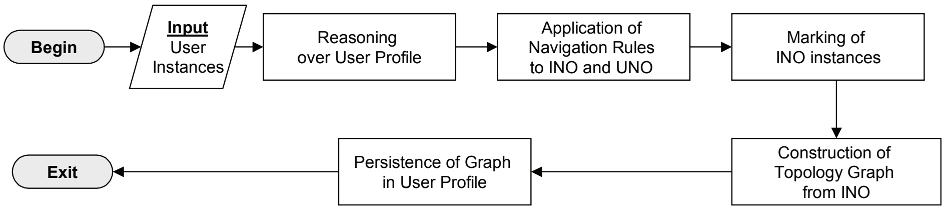

Begin
Input User Profile
Input Navigation Origin and Destination
Unknown User?
YES
NO
Creation of User Instances
Instantiation of Navigation Service
Retrieval of User Profile and relevant Data
Execution of k-shortest paths Algorithm
Computation of Shortest Paths' Quality Score
Path Ranking
Output Shortest and Most Appropriate Path
End

Figure 5. The workflow for the creation of a new navigation service instance


Begin
Input
User
Instances
Reasoning over User Profile
Application of Navigation Rules to INO and UNO
Marking of INO instances
Construction of Topology Graph from INO
Persistence of Graph in User Profile
Exit

is instantiated in the UNO ontology and a new Navigation Service instance is created. Such an instance mainly consists of the topology-graph associated with the specific user. It should be reminded that this graph is created from the INO instances, after the application of the physical navigation rules. Hence, the graph contains all the path elements that are accessible by that user. Such graph is the basic input to the next task, as illustrated in Figure 4, where the *k-shortest paths* between the user and the destination location are computed. Subsequently, for each path, a total *quality score* is calculated, denoting to which extend a path satisfies the perceptual navigation and preference rules. The path with the highest score is the final selected path that is proposed to the user.

The task that refers to the creation of a new navigation service instance is very important, since it involves the most algorithmic and computational parts of the whole system functionality. In Figure 5, one can see that the UNO instances (i.e., user profile) are the main inputs to that task. Those instances pass through a reasoning engine, which results in the user classification with respect to the UNO classes. Subsequently, all types of rules are applied to the INO and UNO knowledge bases. Specifically, the physical navigation rules "mark" the INO instances for further exclusion from the process, whilst the perceptual navigation rules and navigation preferences reward or penalize certain INO instances pertaining to the specific user. The unmarked elements are used for the creation of the user-accessible topology graph, which is also stored in the user profile for future use.

FUTURE TRENDS

The introduction of knowledge engineering technologies in traditional services and applications is expected in the following years. This "paradigm shift" in system design and implementation is merely "pushed" by initiatives like inclusive design and universal access (Stephanidis & Savidis, 2001). However, there are still a lot of open issues before such approach can be massively adopted in commercial systems. One of the greatest challenges is the common definition and adoption of semantic application domain models (i.e., ontologies). In our case, INO and UNO ontologies should have been standardized in some way, so that LBS providers could rely on their specifications in order to develop interoperable semantic navigation services. Moreover, ontological engineering is a very difficult and time-consuming process. Towards simplifying it, there is a great deal of past and current research in developing methodologies and tools for creating, managing, merging, and, updating ontologies (Gomez-Perez, Fernandez-Lopez, & Corcho, 2004). In addition, more and more research projects have commenced to design ontologies for specific application domains. Their attempts, if coordinated accordingly, can result in the creation of an extensible "ontology repository" usable by anyone.

The proposed approach of designing navigation services directly involves human factors. Hence, another issue is the human evaluation of such systems, apart from their performance evaluation. Since semantic LBS are (according to their definition in this article) human-centered, user acceptance and quality assessment are considered as prerequisites before launching them in real-world systems. The parameters that affect the path-selection process should be adjusted carefully by real users prior to system deployment. In the future, we expect to see more research in user evaluation of innovative personalized mobile services. Pervasive computing research has already paved the way for such evaluation frameworks (Scholtz & Consolvo, 2004), but still much progress has to be achieved.

CONCLUSION

This article discusses spatial modeling and processing issues regarding a human-centered navigation service. Such service is mainly targeted to people with navigational limitations, and pursues the vision of context-aware services for ubiquitous computing environments. The main goal of this work is to creatively integrate semantic knowledge engineering technologies with traditional location-based services. In our view, such integration is considered a key enabler for next-generation mobile services that focus on providing advanced user experience.

ACKNOWLEDGMENTS

This work has been partially funded by the Greek General Secretariat for Research and Technology (GSRT) under grant PENED2003 (No. 03ED173).

REFERENCES

Abowd, G. D., Atkeson, C. G., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A mobile context-aware tour guide. *Baltzer/ACM Wireless Networks*, *3*(5), 421-433.

Berners-Lee, T., Hendler, J., & Lassila, O., (2001, May). The semantic Web. *Scientific American, 284*(5), 28-37.

Brachman, R., & Levesque, H. (2004). *Knowledge representation and reasoning*. San Francisco: Morgan Kaufmann.

Duckham, M., & Kulik, L. (2003) Simplest paths: Automated route selection for navigation. In *The Proceedings of COSIT 2003* (LNCS 2825, pp.169-185).

European Institute for Design and Disability. (n.d.). Retrieved from http://www.design-for-all.org/.

Gartner, G., Frank, A., & Retscher G. (2004). Pedestrian navigation system in mixed indoor/outdoor environment: The NAVIO project. *CORP 2004 Geomultimedia04*, Vienna, Austria.

Gomez-Perez, A., Fernandez-Lopez, M., & Corcho, M. (2004). *Ontological engineering: With examples from the areas of knowledge management, e-commerce and the semantic Web*. London: Springer-Verlag.

Grum, E. (2005). *Danger of getting lost: Optimize a path to minimize risk*. Tenth International Conference on Urban Planning & Regional Development in the Information Society (CORP), Vienna, Austria.

Hightower, J., & Borriello, G. (2001). Location systems for ubiquitous computing, *IEEE Computer*, *34*(8), 57-66.

Horrocks, I., Patel-Schneider, P., Harold, B., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web Rule Language combining OWL and RuleML*. World Wide Web Consortium Member Submission. Retrieved from http://www.w3.org/Submission/SWRL/

Hu, H., & Lee, D.L. (2004). Semantic location modeling for location navigation in mobile environment. *IEEE Mobile Data Management*, 52-61.

Leonhardt, U. (1998). *Supporting location-awareness in open distributed systems* (PhD Thesis). London: Department of Computing, Imperial College.

McGuinness, D. L., & Harmelen, F. (2004). *OWL Web ontology language overview*. World Wide Web Consortium Recommendation. Retrieved from http://www.w3.org/TR/owl-features/

Schiller, J., & Voisard, A. (2004). *Location-based services*. San Francisco: Morgan Kauffman.

Scholtz, J., & Consolvo, S. (2004). Toward a framework for evaluating ubiquitous computing applications. *IEEE Pervasive Computing*, *3*(2), 82-88.

Stephanidis, C., & Savidis, A. (2001). Universal access in the information society: Methods, tools, and interaction technologies. *Universal Access in the Information Society*, *1*(1), 40-55.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *IEEE transactions on data and knowledge engineering, 25*(1-2), 161-197.

Wu, Q., & Hartley, J. (2004). Using k-shortest paths algorithms to accommodate user preferences in the optimization of public transport travel. In *Proceeding of UKSIM 2004* (pp. 113-117).

Yen, J. (1971). Finding the k shortest loop-less paths in a network. *Management Science*, *17*, 712-716.

KEY TERMS

Design for All: A design approach that aims at constructing products and services in a way that no user is excluded from using them, independently of her capabilities or limitations.

Indoor Positioning: The determination of an (moving) object's location in an indoor environment.

k Shortest Paths Problem: The identification of a set of paths $\{p_1, \ldots, p_k\}$ between two endpoints s and t (origin and destination, respectively) that satisfy the following criterion: $\text{length}(p_n\text{-}1) \leq \text{length}(p_n)$, for every $n \leq k$, and p_1 is the shortest path between s and t, as computed by a traditional shortest path search algorithm (e.g., Dijkstra).

Navigation: Q service that finds a path between two locations (origin and destination) and gives appropriate instructions in order for the user to successfully follow it and reach the desired destination.

Ontology: A model representing the main entities and their relationships within a domain of discourse. Although similar to other modeling formalisms, ontologies (and especially those based on subsets of logic) can be more expressive and can represent complex restrictions and axioms that govern the domain entities.

Reasoning: The computational procedure that infers new knowledge from explicitly asserted knowledge (expressed through statements and/or rules). Such new knowledge may be of the form of new statement assertions or may just give extra information about the consistency and validity of the existing statements.

Semantic Web: the evolution of the current WWW in a way that it is also machine-understandable in addition to being human-understandable. This evolution is based on the annotation of data with explicit semantics (i.e., metadata), which can describe purpose and attributes and classify it according to some knowledge models (i.e., ontologies).

Sensor Data Fusion for Location Awareness

Odysseas Sekkas
University of Athens, Greece

Stathes Hadjiefthymiades
University of Athens, Greece

Evangelos Zervas
Tei-Athens, Greece

INTRODUCTION

In pervasive computing environments, location is essential information as it is an important part of the user's context. Applications can exploit this information for adapting their behavior. Such applications are termed location-aware applications (e.g., friend-finder, asset tracking). The location of a user is derived by various positioning methods. Especially for indoor positioning, different approaches have been proposed. The majority of indoor positioning systems rely on different technologies, usually of the same kind, like wireless LAN signal strength measurements (Bahl & Padmanabhan, 2000), IR beacons (Sonnenblick, 1998), or ultrasonic signals.

At this point we will quote the definitions of accuracy and precision, the most important characteristics of a positioning system:

- Accuracy denotes the distance within which the system has the ability to locate a user, for example, 1-10 meters.
- Precision denotes the percentage of time the system provides a specific accuracy, for example, 80% of the time the system provides accuracy 1-5 meters (or else accuracy less than 5 meters).

The accuracy and precision are tradable, and it is clear that if we need less accuracy, the precision that the system provides increases.

During the past few years, several location systems have been proposed that use multiple technologies simultaneously in order to locate a user. One such system is described in this article. It relies on multiple sensor readings from Wi-Fi access points, IR beacons, RFID tags, and so forth to estimate the location of a user. This technique is known better as *sensor information fusion,* which aims to improve accuracy and precision by integrating heterogeneous sensor observations. The proposed location system uses a fusion engine that is based on dynamic Bayesian networks (DBNs), thus substantially improving the accuracy and precision.

BACKROUND

Indoor positioning systems have been an active research area since the Active Badge Project (Want, Hopper, Falcao, & Gibbons, 1992). Since then, several indoor location systems have been proposed. A large number of them use IEEE 802.11 (Wi-Fi) access points to estimate location. RADAR (Bahl & Padmanabhan, 2000) is a radio-frequency-based system for locating users inside buildings. It operates by recording and processing received signal strength (RSS) information. The RSS method is used also by the commercial system Ekahau (Ekahau Positioning Engine).

The Cricket Location Support System (Nissanka, Priyantha, & Balakrishnan, 2000) and Active Bat location system (Harter, Hopper, Steggles, Ward, & Webster, 1999) are two systems that use the ultrasonic technology. Such systems use an ultrasound time-of-flight measurement technique to provide location information. They provide accurate location information, but also have several drawbacks like poor scaling and a high installation and maintenance cost. For these reasons they are rather inaccessible to the majority of users.

Another category of location systems uses multiple sensor readings (Wi-Fi access points, RFIDs) and sensor fusion techniques to estimate the location of a user. Location Stack (Graumann, Lara, Hightower, & Borriello, 2003) employs such techniques to fuse readings from multiple sensors. Another similar approach is described in King, Kopf, and Effelsberg (2005). The drawback of these systems is their inability of supporting mobile devices with limited capabilities (CPU, memory) as the location estimation is performed at the client side; hence devices incur the cost of complex computations.

The location estimation system described in this article relies on data from sensors to determine the location of a user. Our work differs from previous approaches in various aspects. Firstly, we use dynamic Bayesian networks for location inference. By using DBNs, we obtain better location estimation results. Along with heterogeneous sensor data that are processed in real time, we can also "fuse"

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

past information about the user. Secondly, our system can support a variety of mobile devices (PDAs, palmtops) with low computing power. Location estimation takes place in a server residing in the fixed network infrastructure. Mobile devices are just transmitting observations from sensors to this server and receive the location estimations. Finally, the adopted system architecture has the advantage of easy management and scalability (e.g., the installation of a new access point is completely transparent to users).

POSITIONING TECHNOLOGIES

In this section we present the principal technologies that are used for indoor positioning and describe their characteristics. We also discuss a categorization of the devices.

The most important wireless LAN standard today is the IEEE 802.11 (Wi-Fi) that operates in the 2.4 GHz ISM band or 5GHz band. This technology is used by several positioning systems that measure the signal strength from access points (RSS) to locate a user.

Radio frequency identification (RFID) is the technology used for security tags in shops, ID cards, and so forth. Tags are powered by the magnetic field generated by a reader and transmit their ID or other information. Such tags do not require any battery and can be deployed in a building to detect object and person passing or proximity.

Infrared (IR) beacons are programmable devices that periodically emit their unique ID in the IR spectrum. Usually the range of these beacons is approximately 10-20 meters, and the infrared receiver should have line of sight with the beacon in order to receive its ID.

Ultrasonic signals are vibrations at a frequency greater than 20 kHz. The devices used to receive and transmit ultrasonic signals are called transducers and are commonly used for distance measuring. In general, they integrate a sensor that can receive or transmit an ultrasonic signal and another RF transmitter/receiver which is used for synchronization.

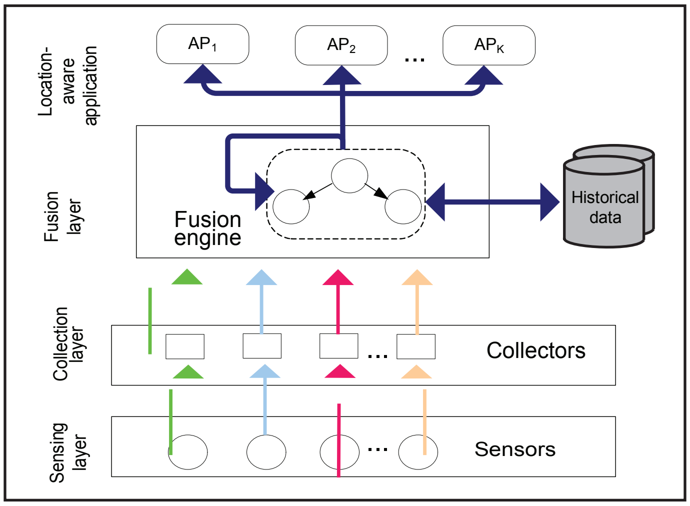
All the previously mentioned devices (elements) of different technologies (access points, beacons, tags, etc.) can be found in indoor environments either deployed in the building or attached to mobile devices. Some of them emit information, and others detect (read) information. According to their position and functionality, the elements can be categorized as follows:

- Portable elements are those carried by users or attached to their mobile devices (RFID tags, Wi-Fi adapters)
- Infrastructure elements are those attached to the building (Wi-Fi access points, IR beacons, RFID tag readers).
- Active elements (sensors) are those which detect a phenomenon or take measurements (RFID tag readers, Wi-Fi adapters).
- Passive elements are those that emit information which is detected by active elements. Wi-Fi access points, IR beacons, and so forth fall into this category.

SYSTEM ARCHITECTURE

The architecture of the proposed location estimation system is organized into three layers: the sensor layer, the collection layer, and the fusion layer. Figure 1 illustrates the generic architecture of the proposed system. In the same figure are also depicted location-aware applications which exploit the location information and databases where the personal profile of users and historical data about their behavior are stored.

Figure 1. Architecture of the indoor location estimation system



The layered approach aims to facilitate effortless inclusion of new elements in order to improve the accuracy and the precision offered by the system. In the following paragraphs we provide a more detailed presentation of each layer.

Sensing Layer

This is the lowest layer of the architecture, comprising sensors of different technologies. Sensors are attached either to the user's mobile device (portable active elements) or to the building (infrastructure active elements). Below, we briefly discuss these two categories.

Portable Sensors

A Wi-Fi adapter can measure the received signal strength (RSS) from a Wi-Fi access point (passive infrastructure element). Similarly, the IR port of a handheld device or a laptop is used as a reader for infrared transmissions from IR beacons that are wall-mounted.

Infrastructure Sensors

RFID tag readers belong in this category. Such readers detect an RFID tag and read its ID when the latter is in proximity. Users can carry RFID tags which have unique IDs. Furthermore, ultrasonic devices, which estimate the distance of a user from a known point, also belong in this category.

Collection Layer

This layer consists of software components called collectors. The role of a collector is to interact with the appropriate sensor and collect measurements or events. Sensors may produce raw data in a variety of formats according to their type. Hence, the output of a Wi-Fi adapter is a stream consisting of RSS measurements from access points; IR and RFID readers generate a stream of proximity events. When such raw data arrive at the collection layer, a preprocessing procedure is performed as described next.

Preprocessing of Raw Data

Assume that a new RSS measurement arrives from a Wi-Fi adapter. Then, the appropriate collector (Wi-Fi collector) quantizes this on *N* discrete levels (values): *S1, S2…SN*. If, for example, the value from the access point with ID *AP2* is between -70 dBm and -60 dBm, the value "*S1*" is assigned to this infrastructure passive element. It should be noted that the number *N* of quantization levels depends mainly on the thresholds (lower and higher) of the access point's transmitted power and the environmental conditions (noise, etc.).

An IR beacon collector, during this preprocessing procedure, operates differently. The two possible states of an IR beacon are: Visible and Not_Visible. Assume that an IR receiver (portable active element) is in the range of the IR beacon with ID IRB3 (infrastructure passive element). This situation will cause a proximity event which will be detected, and thus the collector sets the IRB3 to the value "Visible." An RFID tag reader collector's functionality is similar to the IR beacon collector's, as RFID tag readers detect proximity events, too.

Tuple Forming

After the preprocessing of raw data from the sensing layer, each collector forms a tuple of the type:

(*user_ID, IE_ID, value*)

where *user_ID* is the unique identifier of a user, *IE_ID* is the unique identifier of an infrastructure element, and value is a measurement or an event.

A Wi-Fi collector may form the following tuple:

(userA, AP1, S1)

which denotes that the Wi-Fi adapter (portable active element) of the mobile device of *userA* measures the RSS from access point *AP1* (infrastructure passive element), and the (quantized) RSS has value *S1*.

A possible tuple generated by a RFID tag reader collector would be:

(userB, RFR1, Visible)

which denotes that an RFID tag (portable passive element) worn by *userB* (or attached to his/her mobile device) is in proximity of RFID tag reader with ID *RFR1*.

These tuples are then forwarded to the upper layer where a location estimation procedure is invoked for each user. In the next section, we will show how such values are exploited for location estimations.

Fusion Layer

As mentioned in the introduction, the fusion engine is based on a dynamic Bayesian network, which is used for location inference. Below, we briefly discuss the basic concepts of Bayesian and dynamic Bayesian networks and discuss the adoption of DBNs in the proposed system. We assume that the reader is familiar with the theory of Bayesian and dynamic Bayesian networks. For a more complete introduction, the reader is referred to Jensen (1996) and Mihajlovic and Petkovic (2001).

Figure 2. (a) A Bayesian network (BN) showing four random variables and their dependencies. (b) A dynamic Bayesian network (DBN) showing dependencies between variables in different time-slots.

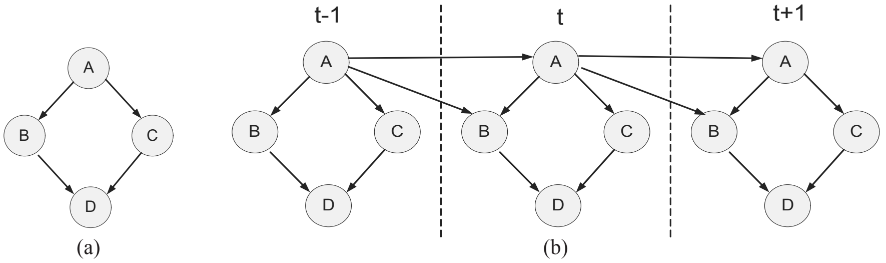
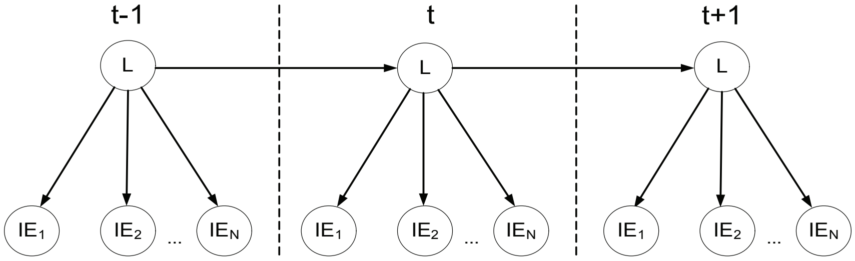


Figure 3. DBN for location estimation representing the dependencies between random variables at different time slots



Bayesian and Dynamic Bayesian Networks

Bayesian networks (BNs) present a statistical tool that has become popular in the areas of machine learning. They are well suited for inference because of their ability to model causal influence (cause-effect) between random variables.

A BN (Figure 2(a)) consists of two parts. The first part is a directed acyclic graph (DAG), representing random variables as nodes and relationships between variables as arcs between the nodes. If there is an arc from a node *A* to a node *B*, it is considered that *B* is directly affected by *A* (*A* is the parent of *B*). Each node is conditionally independent from any other node given its parents.

The second part of a BN is a probability distribution associated with each graph node. This describes the probability of all possible outcomes of the variable given all possible values of its parents. The parameters of this probability distribution would be estimated using observed data (Heckerman, 1995). The DAG and probability distributions together define the joint probability distribution.

A DBN extends the static BN by modeling changes of stochastic variables over time. Random variables in a DBN are also affected by variables from previous time slots (see Figure 2(b)). For simplicity, it is assumed that the parents of a node are in the same or in the previous time slot (First Order Markov Chain).

DBN Integration in the Location System

The DBN that is used in our location estimation system is depicted in Figure 3.

The random variables are:

- the location *L* of the user, which may take values from a set of *K* locations *{L1, L2,…LK}*;
- the *N* infrastructure elements *IE1, IE2,…IEN*. The range of values of those random variables depends on the type of element. Hence, an access point may take a value from the set *{S1, S2,…}* and an RFID tag reader from the set *{Not_Visible, Visible}*.as

The random variable *L* at time *t*, $L^{(t)}$, is directly affected by the random variable *L* at time *t-1*, $L^{(t-1)}$, so $L^{(t-1)}$ is the cause and $L^{(t)}$ is the effect. This is a reasonable assumption as the location of a user is dependent on his/her previous location. Also, the infrastructure elements at time *t* are affected by location at time *t*, $L^{(t)}$; the location of the user affects the value of an infrastructure element (e.g., the signal strength measured from a Wi-Fi access point depends on the location of the user).

The probability distributions that are associated with each node of the DBN are estimated with Bayesian Network learning techniques. In particular, for every infrastructure element *(IE1, IE2,…IEN)*, we estimate the probability distribution $P(IE_i \mid L)$. This can be achieved by taking into account the fixed positions of infrastructure elements, the indoor propagation models of RF and IR signals, the time of flight of ultrasonic signals, and so forth. A simpler technique of learning that that we have adopted for our system is the method of sampling (signal, events) at every location for determining the values of infrastructure elements and the frequency of appearance of these values. According to this frequency we are able to form the probability distributions. In Table 1 we present a probability distribution of a passive infrastructure element (Wi-Fi access point) with ID *AP1*.

Furthermore, the probability distributions $P(L^t \mid L^{t-1})$ for location transition can be generated according to the structure of the building, the distance between two locations, and the time required by a mobile user to cover this distance (see Table 2). It is important to note here that the determination of probability distributions takes place once, at system initialization (training phase).

Location Inference Queries

After having structured the DBN of the fusion engine, we can use it for location estimations. A location inference query might be: "*Where is user X given his/her previous location and given the values (observations) of infrastructure elements associated with this user?*" To answer this, we calculate for each of the K locations *{L1, L2,…LK }* the following conditional probability:

$$P(L^{(t)} \mid L^{(t-1)}, O^{(t)}) \tag{1}$$

which is the mathematical representation of the location inference query and denotes the probability of being at location $L^{(t)}$ at time t (the requested location) given the already known value of the previous location $L^{(t-1)}$ and given the values of the N infrastructure elements at time t, $O^{(t)}$. For simplicity reasons we write:

$$\{IE_1^{(t)}, IE_2^{(t)}, \ldots IE_N^{(t)}\} = O^{(t)} \tag{2}$$

Equation (1) can be converted to the following equation:

$$P(L^{(t)} \mid L^{(t-1)}, O^{(t)}) = \frac{P(L^{(t)}, L^{(t-1)}, O^{(t)})}{P(L^{(t-1)}, O^{(t)})} \tag{3}$$

Taking into consideration that each node of our DBN is conditionally independent from any other node given its parents, we can compute the joint probability that appears in the numerator of (3). Also, as the denominator of (3) does not depend on the random variable $L^{(t)}$, it can be treated as a normalizing constant. Hence, the following equation is derived:

$$P(L^{(t)} \mid L^{(t-1)}, O^{(t)}) = \frac{P(L^{(t)} \mid L^{(t-1)}) * P(O^{(t)} \mid L^{(t)})}{\sum_{i=1}^{K} P(L_i^{(t)} \mid L^{(t-1)}) * P(O^{(t)} \mid L_i^{(t)})}. \tag{4}$$

The probability distributions $P(IE_i \mid L)$ and $P(L^{(t)} \mid L^{(t-1)})$ are known from the training phase, so we can now compute the probabilities for each location {L1, L2,…LK}. Whenever a probability on the numerator of (4) is equal to zero, it is unnecessary to compute the final probability, as it is equal to zero, too. Thus, the calculations are pruned and the overall computation is optimized. The problem of location estimation is to find the location L_i that maximizes the probability.

$$\max\{P(L_i^{(t)} \mid L^{(t-1)}, O^{(t)})\} \tag{5}$$

The location with maximum probability is stored in the database and the profile of the user is updated. Moreover, the location information is forwarded to location-aware applications for the provision of LBS services. After that the system proceeds to the next location estimation in light of the current observations and the previous estimated location of the user.

Table 1. A possible probability distribution for access point with identifier AP1. It can be shown that the probability P(AP1=S2 | L=L1) = 0.3.

	L1	*L2*	…
S1	**0.5**	**0.0**	…
S2	**0.3**	**0.8**	…
…	…	…	…

Table 2. A possible probability distribution for location transition. The probability of transition from L1 to L2 is $P(L^t{=}L2 \mid L^{t-1}{=}L1){=}0.1$.

	L1	*L2*	…
L1	**0.5**	**0.0**	…
L2	**0.1**	**0.8**	…
…	…	…	…

FUTURE TRENDS

Currently we are working on two issues which will have a direct impact on a system's performance and scalability. The first issue is the adoption of a distributed architecture for the system. In this distributed architecture, the building is divided into regions. For each region there is one server responsible for location estimations. Servers of adjacent regions are interconnected in order to interchange information about the users (handovers between regions, etc). The distributed approach of the system will enhance its performance, improve its scalability, and make it more robust in case of server failures.

The second issue that we are working on is the use of "dead reckoning" techniques to improve the precision and accuracy that the system provides. A user's mobile device, which is equipped with an electronic compass and an accelerometer, could provide information about the direction and speed of its owner. Taking also into account the last known position of the user and the time elapsed since then, we can predict the current position and make more accurate estimations.

CONCLUSION

In this article we presented a layered fusion system architecture that exploits information from sensors of different technologies to estimate the location of a user. A key difference from similar systems is the use of dynamic Bayesian networks for location inference. The use of DBNs improves our estimations. Along with sensor information, we take into consideration the previous location of the user thus improving the performance. Additionally, the system supports a variety of mobile devices, including those with restricted computational capabilities (PDAs, etc.) as they do not incur the burden of complex location calculations.

ACKNOWLEDGMENTS

This work was performed in the context of the "PENED" Program, co-funded by the European Union and the Hellenic Ministry of Development, General Secretariat for Research and Technology (Research Grant 03ED173).

REFERENCES

Bahl, P., & Padmanabhan, V. (2000). RADAR: An in-building RF-based user location and tracking system. *Proceedings of IEEE INFOCOM* (pp. 775-784). Tel-Aviv, Israel.

Ekahau. (n.d.). *Ekahau positioning engine.* Retrieved from http://www.ekahau.com/

Graumann, D., Lara, W., Hightower, J., & Borriello, G. (2003). Real-world implementation of the location stack: The universal location framework. *Proceedings of the 5th IEEE Workshop on Mobile Computing Systems & Applications (WMCSA 2003)* (pp. 122-128).

Harter, A., Hopper, A., Steggles, P., Ward, A., & Webster, P. (1999). The anatomy of a context-aware application. *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (Mobicom '99).*

Heckerman, D. (1995). *A tutorial on learning with Bayesian networks.* Technical Report MSR-TR-95-06, Microsoft Research, USA.

Jensen, F. (1996). *An introduction to Bayesian networks.* New York: Springer-Verlag.

King, T., Kopf, S., & Effelsberg, W. (2005). A location system based on sensor fusion: Research areas and software architecture. *Proceedings of the 2nd GI/ITG KuVS Fachgespräch "Ortsbezogene Anwendungen und Dienste,"* Stuttgart, Germany.

Mihajlovic, V., & Petkovic, M. (2001). *Dynamic Bayesian networks: A state of the art.* Technical Report, Center for Telematics and Information Technology, University of Twente, The Netherlands.

Nissanka, B., Priyantha, A., & Balakrishnan, H. (2000). The cricket location-support system. *Proceedings of MOBICOM 2000* (pp. 32-43). Boston: ACM Press.

Sonnenblick, Y. (1998). An indoor navigation system for blind individuals. In CSUN Center On Disabilities (Ed.), *Proceedings of the CSUN 1998 Conference,* Los Angeles, CA.

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The Active Badge location system. *ACM Transactions on Information Systems, 10,* 91-102.

KEY TERMS

Active Element (Sensor): One of the elements that detect a phenomenon or take measurements, like RFID tag readers and Wi-Fi adapters.

Bayesian Network (BN): A directed acyclic graph where nodes represent random (stochastic) variables, and arcs represent dependence relations among these variables.

Data Fusion: The combination of data derived from heterogeneous sources such that the resulting information is better than it would be if these sources were used individually.

Dynamic Bayesian Network (DBN): The extension of static Bayesian network by modeling changes of stochastic variables over time.

Infrastructure Element: One of the elements that are attached to the building, like Wi-Fi access points, IR beacons, and so forth.

Pervasive Computing: The integration of computation into the environment in order to offer a broad range of services to users.

Positioning: The capability to detect the location of a wireless device carried by a user (e.g., cell phone).

Service Delivery Platforms in Mobile Convergence

Christopher J. Pavlovski
IBM Corporation, Australia

Laurence Plant
IBM Corporation, Australia

INTRODUCTION

The demand for enriched multimedia content and entertainment services in mobile networks is being largely driven by the emergence of mobile broadband. A key problem for institutions attempting to capitalize on these new channels for service delivery is a capability to deploy many multimedia services rapidly and cost effectively. Traditional approaches in deploying new services have largely focused on discrete systems for each new service, often termed point solutions or silos. Recent emerging standards coupled with implementation constraints have led to the development of a more strategic approach. Such an approach involves the creation of a service delivery platform (SDP), capable of delivering a broad range of content and services from a host of multimedia applications. Several initiatives are attempting to lay the foundations for the architecture and framework for SDP solutions that support the emerging multimedia and entertainment services for mobile devices. Recent initiatives include platforms based upon the IP multimedia system (IMS), Parlay X, and IT standards-based designs.

A central characteristic of the SDP approach to mobile service delivery is the capability to supply numerous services to mobile users with observed reductions in elapsed effort to bring these services online; this also bestows cost reduction and speed to market. The benefits of a service delivery business model are applicable to the mobile operator, mobile customers, and external third-party developers. Customers are offered more services quickly, while third-party developers are able to focus on core capabilities of their intended service, collectively offering benefits in terms of time to market and reduced cost. The business benefits illustrate why this service delivery approach is recently gaining increased attention by mobile operators globally.

In this article we outline the fundamental principles of a service delivery platform and the business model to be addressed in mobile convergence. The emerging standards and reference architectures are presented, and their shortcomings are discussed. We outline the key requirements that a service delivery platform is expected to address from an operator perspective and summarize the key technology design points. We also outline several future trends in how this emerging mobile technology is being deployed in various application scenarios. This involves straightforward mobile news services, through gaming, and complex interactive multimedia scenarios for the mobile device. These new services make further convergent demands upon three technology domains: the mobile network, IT systems, and the content/media sources.

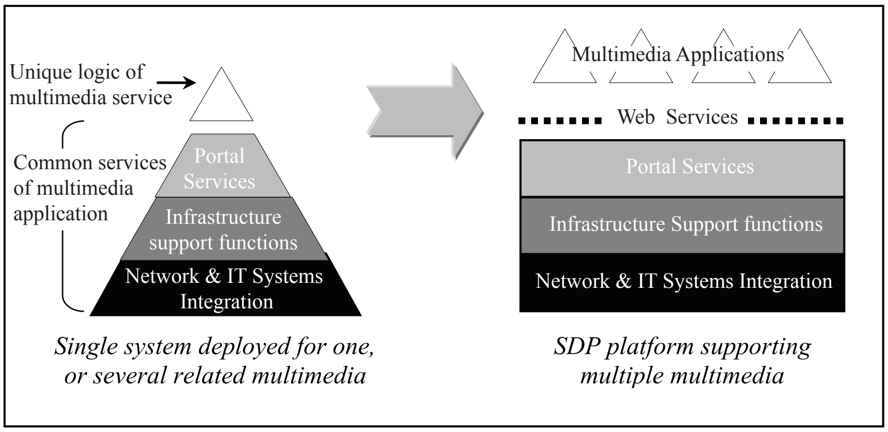
BACKGROUND

Traditional approaches to deploying new multimedia services involve development of a discrete system to deliver one or a related set of services. This approach involves the development of several common functions required by the multimedia service. A recent trend by many operators globally is deployment of one common service delivery platform that supports multiple applications. The SDP is intended to contain all the common functions and services that a wide range of applications may require in order to deliver its service or function. Figure 1 illustrates the change in design philosophy (Pavlovski & Staes-Polet, 2005), where traditional deployment involves development of common delivery functions for each (or a related set of) multimedia service(s). The SDP approach transforms this by combining the common functions used by multimedia services into one platform that may be exploited by a range of multimedia applications.

Generally, the term convergence*,* when used in the context of mobile networks, is used to denote the rationalization of internetworking technologies and protocols. An additional form of convergence is related to the convergence of several operational and business domains—more specifically, the convergence of information technology systems, the networks, and media/content applications. Such integration imposes additional complexity, and the service delivery platform is ideally suited to address this type of convergence. This notion of convergence and applicability of the service delivery platform is recently gaining widespread attention, with several aspects of these emerging service delivery

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Transformation to service delivery platform


Unique logic of multimedia service
Common services of multimedia application
Portal Services
Infrastructure support functions
Network & IT Systems Integration
Single system deployed for one, or several related multimedia
Multimedia Applications
Web Services
Portal Services
Infrastructure Support functions
Network & IT Systems Integration
SDP platform supporting multiple multimedia

platforms actively studied within research and industry (Hanrahan, 2006; Deckers, 2006; Kimbler, Stromberg, & Dyst Appium, 2006).

Mobile devices such as cellular phones, portable digital assistants, and tablets are becoming increasingly adorned with new services and media format. The fundamental business problem is to successfully integrate the network, information technology, and content applications in a unified manner that ameliorates costs for mobile operators, while supporting rapid deployment of new services in a cost-competitive manner.

SERVICE DELIVERY PLATFORMS

The method used to construct service delivery platforms is based on either a network-centric or an information technology centric (IT-centric) view of the problem domain. Platforms based on the Parlay X or IMS standards and frameworks may be categorized as network-centric; there are several notable examples (Pailer, Stadler, & Miladinovic, 2003; Akkawi, Schaller, Wellnitz, & Wolf, 2004, Magedanz, Witaszek, & Knuttel, 2005; Hanrahan, 2006). In contrast, several broader attempts have been discussed in the literature that apply an IT-centric design (Pavlovski & Staes-Polet, 2005; Hwang, Park, & Jung, 2004; Pavlovski, 2002b). These platform design styles largely reflect the heritage of the originating body. Moreover, IMS and Parlay X have emerged from network standards bodies, while the other works reflect practical IT experiences in building service delivery platforms for mobile operators.

Regardless of the approach taken, the underlying business model remains the same. In this section we first outline this business case, and then present the two strategies in addressing the service delivery model, outlining the benefits and shortcomings of each design viewpoint.

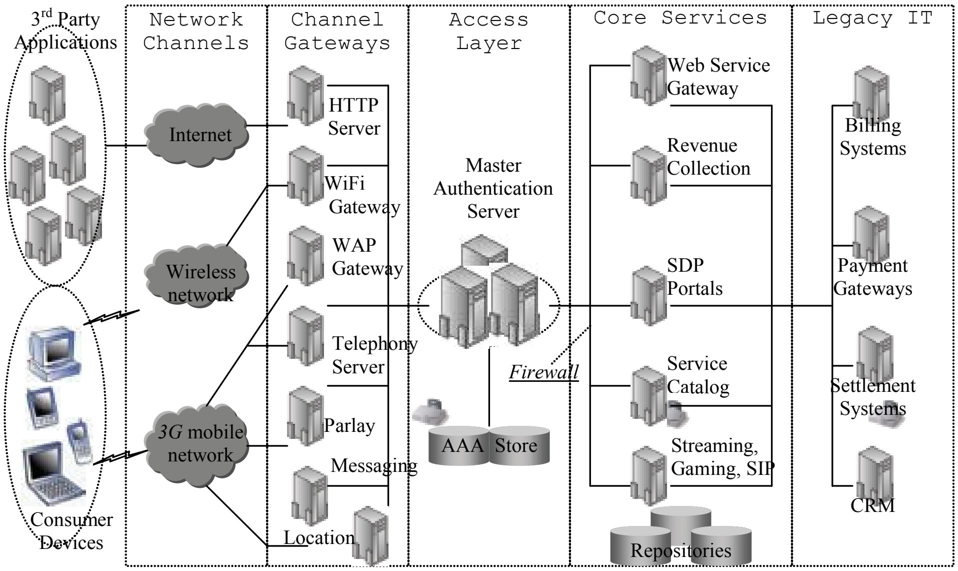
The Service Delivery Business Model

The principle advantage of an SDP for mobile operators is the ability to reduce the cost for deploying new services, while increasing the potential for generating revenue (Deckers, 2006). It is suggested that around one-third of the effort required to introduce a new service is attributed to developing the unique business logic of the service (Pavlovski & Staes-Polet, 2005). This means significant capability is common to a range of services and may be placed into a consolidated platform (see Figure 1). By combining this common functionality within the service delivery platform, the cost for building the platform is amortized since this may now be leveraged by several applications.

Since the SDP is deployed by a mobile operator, external third parties are then able to develop applications that deliver mobile content. The third-party developer may now focus on building the unique business logic associated with the intended multimedia service or content, and is able to reduce the costs by leveraging a set of common capabilities within the service delivery platform.

The service delivery business model for mobile applications has its origins with the *iMode* service (Pavlovski, 2002a), where the platform enables third-party application providers to deliver their content or service, via an iMode service platform, to mobile users. While the technologies have changed considerably over time, the service delivery business model has largely remained constant. The elementary model contains three entities: the mobile network operator, mobile customers, and external third-party developers. The mobile operator (alternatively this may be an MVNO) hosts the SDP and provides a set of common services to the external third-party developer from which to build multimedia applications. The mobile operator owns and maintains the relationship with mobile customers and is able to bring this consumer market to the external third-party developer. A key benefit

Figure 2. IT-centric service delivery platform


3rd Party Applications
Network Channels
Channel Gateways
Access Layer
Core Services
Legacy IT
Internet
Wireless network
3G mobile network
Consumer Devices
HTTP Server
WiFi Gateway
WAP Gateway
Telephony Server
Parlay
Messaging
Location
Master Authentication Server
Firewall
AAA Store
Web Service Gateway
Revenue Collection
SDP Portals
Service Catalog
Streaming, Gaming, SIP
Repositories
Billing Systems
Payment Gateways
Settlement Systems
CRM

of this relationship is that once the platform is constructed, the mobile operator is able to rely on external third parties to build, at their expense, new services. The principle motivation for the external developer, which may be an enterprise, is marketing access to the large customer base owned by the mobile operator. The advantages for the mobile customer is ease of access, a greater range of multimedia and content services, with the payment mechanisms for such access generally viewed as secure (Devine, 2001).

IT-Centric Design Viewpoint

From an IT perspective the key principle is to abstract a range of network and IT services in a way that enables external third-party applications to be developed in a straightforward manner (Pavlovski, 2002a). Such an environment is sometimes referred to as a service creation environment (Schulke et al., 2005). Figure 2 illustrates the topology of an IT-centric SDP.

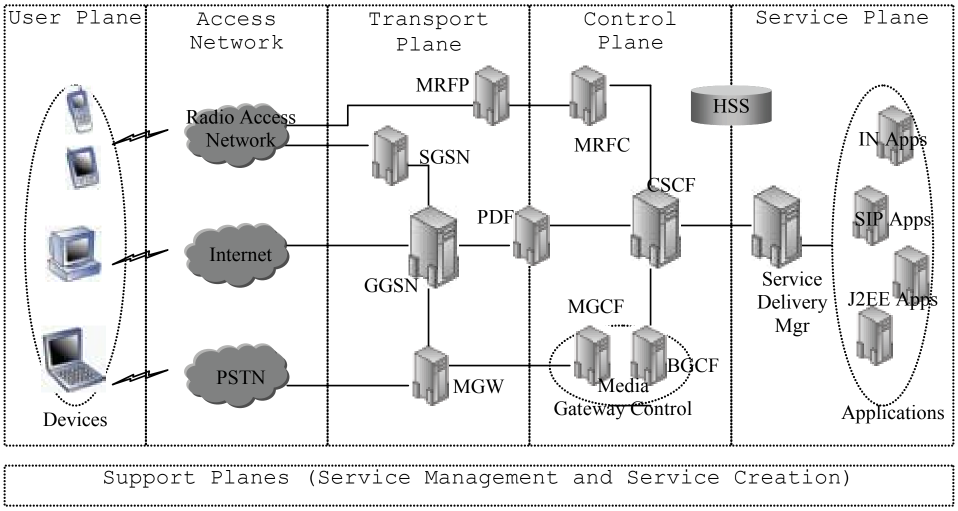
A layered architecture is a key theme of this design, with each layer abstracting the services provided on the preceding layer. This allows the technology to alter without impact to all components within the platform. At the outer most level, the customer and third-party applications reside, external to the platform; note that additional internal applications may also be deployed by the mobile operator. Customers gain access to services provided by multimedia applications via the SDP. The network channel layer defines the various networks used to deliver content and services to consumer devices. The remaining components of an SDP are deployed within the channel gateway, access layer, and core services layer. These layers are now described in more detail.

Channel gateways include those nodes responsible for integration and accessing various mobile and fixed networks. This includes the parlay gateway, messaging gateways (i.e., SMSc and MMSc), WAP gateways, and location and telephony servers. This underpins the IT-centric view of an SDP by viewing access to the network as an abstract service that the SDP builds upon. These network services are made available to application developers via the Web service gateway. Common services include the ability to send and receive messages, call control, initiating multimedia content downloads, and managing delivery receipt notifications for content delivery.

The access layer hosts the master authentication server. As the central security node, this component conducts authentication of customers accessing the SDP platform and also authenticates access by external third-party applications that make use of the published Web services. In order to provide customers with trusted access to third-party applications, a federated identity management scheme is typically employed, for instance as defined by OASIS. This means that a customer need only login once to the SDP and may then access multiple third-party applications without the need to re-authenticate. Authentication exchange occurs on the user's behalf between the SDP and external third-party application, for example using the Security Assertion Markup Language (SAML).

Core services include streaming and gaming servers, digital rights management, a service catalog that provides a menu on the mobile device for each multimedia service available, several portals, and revenue collections facilities such as billing and payment. A key node is the Web service gateway which typically offers a range of network and IT-

Figure 3. Network-centric service delivery platform



related services to third-party developers. This may include sending messages, prepaid, postpaid charging, payment, and additional authorization services. The portals furnish the visual interface accessed by customers, third-party developers, and internal administrators of the platform. Hence, there are generally several portals available to cater to the needs of each user community. The customer portal provides services to conduct registration, subscribe to services, and manage accounts. A service relationship management (SRM) portal is a vital component allowing third-party developers to register new applications, monitor usage, review revenues due, and provide an environment that facilitates development of new services; this may include testing and development tools (i.e., service creation environment).

While clearer separation between the network and SDP environment is beneficial, particularly as network technologies evolve, there are several drawbacks. The design relies upon a mature network with standardized mechanisms for network access and control. Consequently, deployment under this assumption may be problematic where network access is inconsistent. IT-based SDPs generate billable events but do not themselves manage the balance or prevent access to services based on insufficient funds. Rather, they rely upon external systems for balance management. Furthermore, other than security and Web service definition, the supporting standards and literature are limited and emerging.

Network-Centric Design Viewpoint

Service delivery frameworks that have emerged from standards bodies, such as IMS, Parlay X, and multimedia domain (MMD), largely reflect a network-centric design. In Magedanz et al. (2004, 2005), a service delivery platform that extends the IP multimedia system has been developed. The authors outline a test-bed architecture that integrates the session initiation protocol (SIP) and Parlay with the telecommunications network to deliver multimedia services. A further platform for delivery of mobile games over the IMS has also been described (Akkawi et al., 2004). While there appears to be more standards work on the network-centric design, the notion of IT-based systems is observed (Magedanz & Sher, 2006).

In broad terms, IMS specifies an architecture that supports IP telephony and multimedia services such as instant messaging, videoconferencing, voice mail, and multiparty gaming (3GPP, 2001). Communications require session establishment between user devices and application servers, and the signaling protocol selected by IMS to establish these sessions is SIP. The architecture defines a layered model for the telecommunications network referred to as planes, similar in concept to the layers of the IT-centric design. These IMS planes are depicted in Figure 3 and include: *access networks,* being the physical network the end user connects to; *transport plane,* which is the common IP backbone network of the telecommunications service provider; *control plane,* which provides key functions of authentication, session establishment, and quality of service; and *service plane,* which houses application servers and provides an abstracted network interface.

The access network connects the cellular, or fixed, network from the physical internetworking equipment to the customer premise or mobile device. For a device to connect to an IMS network, it needs to support an IMS client, providing various functions including the visual and audio presentation of the service to the user. Packets of data which encapsulate the service, such as VOIP from a PC, IPTV to a set top box, or instant messaging to a mobile device, are all carried on a common transport plane. This plane contains the network elements, the gateway GPRS support node (GGSN) and serving GPRS support node (SGSN), media servers (media resource function processor (MRFP)) for playing announcements, and media gateways (MGW) for interconnecting IP traffic with other networks such as the PSTN.

Control plane authentication of the user's device is achieved by lookup to the home subscriber server (HSS), verifying that the device is enrolled and able to connect to the network and use the requested service. Session establishment is managed by the call session control function (CSCF), which connects the user to the correct application server in the service plane. The policy decision function (PDF) assigns resources to manage quality of service to ensure there is sufficient bandwidth available to deliver the requested service. Several nodes manage the media gateway (MGW) and internetworking between IP and SS7; this includes the media gateway control function (MGCF) and the breakout gateway control function (BGCF), which selects the required (local or foreign) network. As the broker between the MRFP and media applications, the media resource function controller (MRFC) is intended to perform conference, roaming, and media control; however, these capabilities are often developed directly within the applications of the service plane.

The aim of the service plane is to provide an environment for executing various common IP services such as presence, location, messaging, videoconferencing, and multimedia. This is largely where the multimedia and telephony applications reside and is the least defined aspect of the architecture. IMS has to date largely focused upon network convergence, the establishment of IP sessions between users on different networks, how to ensure quality of service, and implementation of heritage features from the mobile phone network such as roaming between networks. As technology migrates to VOIP services, quality of service, mobility, and roaming across different access networks are seen as key features of IMS which will enable the telecommunications service provider to grow the business.

Integral to the network-centric model is near-real-time balance management for the prepaid market, where there needs to be sufficient funds available within the customer account before the control plane grants access to the service being requested. This model becomes complicated when the customer may not select the desired service until the control plane has already allowed access to the service plane, particularly as the customer has access to the shopping portal resident in the service plane. In this instance, like the IT-centric model, the SDP needs to advise the balance manager what service is being requested before the service is initiated.

A drawback of the IMS-based approach is lack of detail for SDP structures such as third-party relationship management, service or device management, and how to deliver services in the service plane. This reflects the heritage of the authoring body, with specifications emerging from network engineering standards bodies. The focus is upon network application service delivery, omitting detail required for broader IT-based multimedia service delivery. Such capability would be added by introducing a service delivery manager (see Figure 3), which contains much of the detail of an IT-centric design.

FUTURE TRENDS

Mobile devices continue to evolve in capability and computational strength. Together with the increasing reach of mobile networks, access to multimedia content and services will become ubiquitous. These factors contribute to trends in offering an increasing number of multimedia services to mobile customers. This includes gaming, podcasting, telephony services, and broadcast and video on demand. Digital rights management is as yet not aggressively implemented. However, as more content moves to this mobile digital environment, effective protection of digital content will become increasingly important.

Perhaps the key consideration in the evolution of service delivery is the convergence of network IT systems and applications. As mobile networks move towards IP-based telephony, the distinction between the network and IT boundary grows dim, placing greater emphasis on well-defined standards and architectures. For instance, scenarios where users are able to view video content on a mobile device while roaming, pausing this when arriving home, with the ability to continue viewing using the home television, exemplify further the prospective convergence between mobile and fixed networks in the delivery of multimedia services.

Services-oriented architecture (SOA) is emerging as the prevalent cross-industry IT integration regime and is likely to gain favor as a means of integrating components of an SDP. Given that integration of network IT and applications is central to an SDP, SOA may provide a means for IT and network-centric SDP models to converge.

CONCLUSION

The standards for service delivery are still emerging, with specifications providing more detail regarding the network aspects of service delivery. An IT-centric SDP design lends more attention to the detail of integration with IT systems and third-party developers. This is beneficial where many implementations underestimate the effort and complexity of integrating with existing legacy systems. While there is a general lack of SDP standards for an IT-centric approach, Web service standards specified by OASIS in identify management and security exchange may be effectively applied. A network-centric design is well defined at the network level, offering a greater range of network services. However, these implementations still require the additional consideration for supporting business processes, particularly for third-party service relationship management and application provisioning.

Either approach largely reflects the heritage of the engineering discipline. However, as both the network and IT systems continue to converge, these boundaries will become less visible. Hence the need for clear architectural separation, as layers, becomes more important to ensure technology independence is maintained. Notwithstanding the design style adopted, there are several key considerations in an SDP to ensure the underlying business model is addressed. This includes the provision of an environment for rapid creation of multimedia applications, an intuitive user portal experience as an interface to a full range of content and services, and a comprehensive portfolio of Web services to enable the development of rich multimedia applications and services by external third parties.

REFERENCES

Akkawi, A., Schaller, S., Wellnitz, O., & Wolf, L. (2004). A mobile gaming platform for the IMS. *Proceedings of the 3rd International Workshop on Network and System Support for Games* (Netgames 2004), Portland, OR.

Anegg, H., Dangl, T., Jank, M. et al. (2004). Multimodal interfaces in mobile devices—the MONA Project. *Proceedings of the Workshop on Emerging Applications for Mobile and Wireless Access (WWW2004)*, New York.

Deckers, G. (2006). Cost down, revenues up: SDP business case. *Proceedings of the 10th International Conference on Intelligence in Service Delivery Networks (ICIN)*, Bordeaux, France.

Devine, A. (2001). *Mobile Internet content providers and their business models.* Masters thesis, Department of Electrical Engineering and Management, The Royal Institute of Technology, Sweden.

Hanrahan, H. (2006). Towards a standards based service delivery platform using service oriented reference points. *Proceedings of the 10th International Conference on Intelligence in Service Delivery Networks (ICIN)*, Bordeaux, France.

Hwang, T., Park, H., & Jung, J. W. (2004). The architecture of the digital home services delivery with OSGi. *Proceedings of the IASTED Conference on Communication Systems and Applications (CSA 2004)*. Banff, Canada.

Kimbler, K., Stromberg, A., & Dyst Appium, J. (2006). The role of convergent service delivery platform in service migration to IMS. *Proceedings of the 10th International Conference on Intelligence in Service Delivery Networks (ICIN)*, Bordeaux, France.

Magedanz, T., & Sher, M. (2006). IT-based open service delivery platforms for mobile networks: From CAMEL to the IP multimedia system. In P. Bellavista & A. Corradi (Eds.), *The handbook of mobile middleware.* Chapman & Hall/CRC Press.

Magedanz, T., Witaszek, D., & Knuttel, K. (2004). Service delivery platform options for next generation networks within the national German 3G beyond testbed. *Proceedings of the South African Telecommunication Networks Architectures Conference (SATNAC04)*, Stellenbosch, South Africa.

Magedanz, T., Witaszek, D., & Knuettel, K. (2005). The IMS Playground @ Fokus—an open testbed for next generation network multimedia services. *Proceedings of the 1st International IEEE Computer Society Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM'05)* (pp. 2-11).

Pailer, R., Stadler, J., & Miladinovic I. (2003). Using Parlay APIs over a SIP system in a distributed service platform for carrier grade multimedia services. *Wireless Networks, 9*(4), 353-363.

Pavlovski, C.J. (2002a). Reference architecture for mobile Internet service platform. *Proceedings of the 2nd Asian International Mobile Computing Conference (AMOC 2002)*, Langkawi, Malaysia.

Pavlovski, C.J. (2002b). Software architecture for mobile Internet service platform. *Proceedings of the Workshop on Pervasive Computing, Going Beyond the Internet for Small Screens (OOPSLA 2002)*, Seattle, WA.

Pavlovski, C.J., & Staes-Polet, Q. (2005). Digital media and entertainment service delivery platform. *Proceedings of the 1st ACM International Workshop on Multimedia Service Composition (MSC '05)* (pp. 47-54). Singapore.

Schulke, A., Kovacs, E., Stuttgen, H., Akkawi, A., Kuhnen, M., Riu, A., & Winkler, F. (2005). Creating new communication services efficiently. *NEC Journal of Advanced Technology, 2*(2), 170-178. Retrieved from http://www.nec.co.jp/techrep/en/r_and_d/a05/a05-no2/a0502p170.html

3GPP (3rd Generation Partnership Project). (2001). *Technical specification group services and system aspects.* Service Requirements for the IP Multimedia Core Network Subsystem (Stage 1) (Release 5) 3G TS 22.228 V5.0.0 (2001-01).

KEY TERMS

Federated Identity Management: A standard that enables a user to use one set of credentials to sign on and access the networks of several enterprises in order to conduct transactions.

Multimedia Domain (MMD): Specified by the 3GPP2 group, a set of specifications for the CDMA network based on IMS and Parlay X.

Mobile Network Virtual Operator (MVNO): Being able to sell branded mobile network services without owning a mobile network, through an established relationship with a mobile network operator.

Organization for the Advancement of Structured Information Standards (OASIS): A not-for-profit, international consortium that drives the development, convergence, and adoption of e-business standards.

OSE (Open Mobile Alliance) Service Environment: A reference architecture for how services may be combined to form new services (Schulke et al., 2005). The standard is intended to promote common architectural principles and broadly defines the functional entities, within a telecommunications service environment, for application developers and service providers.

Parlay X: Web service extension to the Parlay APIs that allow developers to build applications using the Parlay API. The Parlay API is an open interface to the mobile and fixed telephone network.

Security Assertion Markup Language (SAML): An XML-based framework that defines an interoperable standard for exchanging security information.

3rd Generation Partnership Program (3GPP): A collaboration between several telecommunications standards bodies to develop 3G GSM technical specifications.

Service Provision for Pervasive Computing Environments

Emerson Loureiro
Federal University of Campina Grande, Brazil

Frederico Bublitz
Federal University of Campina Grande, Brazil

Loreno Oliveira
Federal University of Campina Grande, Brazil

Nadia Barbosa
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Glauber Ferreira
Federal University of Campina Grande, Brazil

INTRODUCTION

The fast development on microelectronics has promoted the increase on the computational power of hardware components. On the other hand, we are facing a significant improvement on energy consumption as well as the reduction of the physical size of such components. These improvements and the emergence of wireless networking technologies are enabling the development of small and powered mobile devices. Due to this scenario, the so-called pervasive computing paradigm, introduced by Mark Weiser in 1991 (Weiser, 1991) is becoming a reality. Such a paradigm envisions a world where environments are inhabited by computing devices, all of them seamlessly integrated into peoples' lives, and effectively helping to carry on their daily tasks.

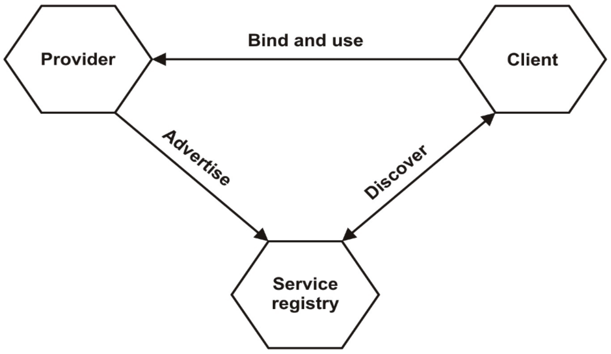
Among others, one major characteristic of Weiser's vision is that each device in an environment becomes a potential client or provider of resources. Not surprisingly, pervasive computing environments are becoming dynamic repositories of computational resources, all of them available to mobile users from the palm of their hands. However, devices can unpredictably join and leave such environments. Thus, resources can be dynamically made available or unavailable. Such a scenario has a great impact on the way that resources are found and used. In the case of static environments, such as the Web, it is reasonable to look up and access resources, such as Web pages, knowing the address of their providers beforehand. On the other hand, for dynamic environments, such as the pervasive computing ones, this is not a reasonable approach. This is due to the fact that one cannot guarantee that the provider of a resource will be available at any moment, because it may have left the environment or simply turned off. A better approach would be to discover these resources based on their descriptions, or any other feature that does not require the client to know the specific address of their providers.

To this end, some of the current pervasive computing solutions, like Wings (Loureiro, Bublitz, Oliveira, Barbosa, Perkusich, Almeida, & Ferreira, 2006), Green (Sivaharan, Blair, & Coulson, 2005), RUNES (Costa, Coulson, Mascolo, Picco, & Zachariadis, 2005), and Scooby (Robinson, Wakeman, & Owen, 2004), are making use of a novel approach from the branch of distributed applications, the *service-oriented computing* paradigm (Papazoglou, 2003; Huhns & Singh, 2005). This is due to the fact that such a paradigm provides a crucial element for pervasive computing systems, the ability for dynamically binding to remote resources (Bellur & Narenda, 2005), which enables mobile devices to find needed services on demand.

However, pervasive environments may be structured in different ways. They can range from wired networks to completely wireless ones, where communication among the

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. General view of a service-oriented architecture



devices is performed in an *ad hoc* way. Such a characteristic indicates that the way services are provisioned in a pervasive computing environment should fit in its organization, in order to enhance the access to the services available.

Considering the above discussion, in this article we provide a review on service provision and its applicability in pervasive computing. More precisely, we will list the existing service provision approaches and discuss the characteristics and problems associated with each one, as well as their usage in pervasive computing environments. We start by providing introductory concepts of service-oriented and pervasive computing, respectively in the service-oriented computing and pervasive computing sections. Next, we present the service provision techniques available and how they can be applied for pervasive computing environments. The main current solutions within this scope will be introduced in the service oriented technologies section. Some of the future trends associated with research for service provision in pervasive computing environments will be presented in the future research trends section. Finally, in the conclusions section we present the conclusions of this article.

SERVICE-ORIENTED COMPUTING

The service-oriented computing (SOC) paradigm has been considered as the next step in distributed computing (Papazoglou, 2003). In a general way, this paradigm can be viewed as the development of applications through the runtime integration of software pieces named of *services* (McGovern, Tyagi, Stevens, & Mathew, 2003). In this process, three elements are involved: defining what is known as a *service-oriented architecture* (SOA), a *service client,* a *service provider*, and a *service registry*. The former is the one who whishes to use a service. Conversely, service providers are those which offer services for potential clients. Finally, the service registry is where providers advertise or announce their services (through service advertisements), enabling clients to dynamically discover them. By dynamic, we mean that clients are capable of discovering services at runtime, thus providing a high degree of flexibility for applications. Once clients have discovered a service, they are able to bind to it; that is, to create a link with the service, in order to use it (through a proxy to the real service). This process of advertising, discovering, binding, and using a service is illustrated in Figure 1.

In open environments, the dynamic discovery of services implies that they can be used by heterogeneous clients. Within this scope, heterogeneity is concerned with features like the operating system running in each client and the hardware platform it has been built on. As a consequence of such heterogeneity, for enabling an application to flexibly integrate services, they should present the following features (Papazoglou, 2003):

- **Loose Coupling:** A service must not require from the clients any knowledge about its internal implementation.
- **Implementation Neutrality:** The usage of services must not rely on any specific programming language, operating system, or hardware platform.
- **Dynamically Discoverable:** Services should be discovered at runtime.

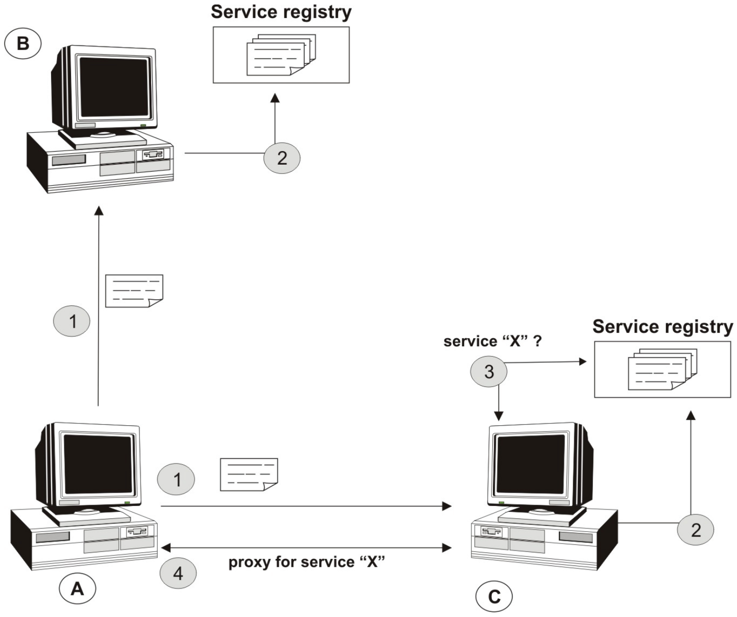
PERVASIVE COMPUTING

The field of pervasive computing has its origins at the Xerox Palo Alto Research Center. The pioneer work that has been led there has culminated in the novel article of Mark Weiser in 1991 (Weiser, 1991), where he describes the fist ideas of pervasive computing. Weiser's vision is at the same time revolutionary and simple: a world where computing is embedded in everyday objects, like cars, televisions, and air conditionings, all seamlessly integrated into our lives and performing tasks for us (Turban, Rainer, & Potter, 2005). When Weiser talked about seamless integration, he meant that applications running in these objects should act proactively on our behalf. They should, for example, present us with relevant information, based on what we want/need and the resources (e.g., a printer) available in the environment we are immersed.

SERVICE PROVISION APPROACHES IN PERVASIVE COMPUTING

When provisioning services in a pervasive environment, one aspect to be considered is the way it is organized; that

Figure 2. Example of a push-based service provision



is, whether the environment is based on a wired network infrastructure, whether it is formed in an *ad hoc* way, or both. This is necessary for dealing with the particularities of each environment, and within this scope, we can say that there are two major ways of performing service provision (Nickull, 2005): the *push-based* and the *pull-based* approach. In the next sections we will outline each of these approaches as well as describe how they fit into pervasive computing environments.

Push-Based Service Provision

In this approach, a provider advertises its services directly to potential clients. In other words, it sends service advertisements to all network hosts, either they are interested on the service or not. Such advertisements, when received by a host, will be kept in a local service registry. Therefore, once a client wants to discover some service, it will then look up in such a registry for the services that match its needs. An example of the push-based service provision is illustrated in Figure 2. Note that the provider, host A, sends the advertisement of a service directly to hosts B and C, as illustrated in Step 1. When this advertisement is received, it will be stored on a local service registry at each host (Step 2). Then, once a host, in our example host C, wants to discover a service, it then inquires this registry, as illustrated in Step 3. Finally, considering that a relevant service has been found, it is possible to ask its provider (host A) for a proxy to the service (Step 4), enabling host C to use it.

Using this approach, one major problem to be pointed out is about the validity of the service advertisements. It is concerned with the fact that a service provider can leave the network, but the advertisements associated with its services can still be available. One approach which could be used for solving this problem is to require the provider to explicitly notify the network hosts about its leaving, and consequently its services will be no longer available. The problem is that it is not always possible to do that. For example, if the provider is a mobile device, it may be suddenly run out of energy. To deal with this, providers could be aware of the energy level of the device, in order to notify the network hosts that within some minutes it may not be accessible anymore. However, other factors can be involved in the leaving of a provider from the network. It can be just turned off by its user, or leave the network coverage area. Keeping track of all these factors is a task that certainly overloads the provider. A more reasonable solution would be requiring both providers and clients to cooperate for renewing the service advertisement. Therefore, as long as the advertisements are renewed, it is possible, but not guaranteed, that the service is available. On the other hand, when the advertisement has not been renewed within a time interval, then the service is probable, but also not guaranteed, to be unavailable, either because its provider has left the network or because the service has been unadvertised.

One interesting point to notice is that the push-based service provision does not require any physical infrastructure. This means that such an approach is well suited in decentralized and/or infrastructure-less pervasive environments. One problem, in the scope of pervasive computing, is that the advertisement task can consume a lot of bandwidth if many devices are provisioning services in the environment. Therefore, in environments with very limited wireless links this is certainly a major problem. On the other hand, as services are searched locally, the discovery process does not involve costs of communication.

Pull-Based Service Provision

In the pull-based approach, in order to discover services clients must inquiry remote registries for the needed services. This can be performed in two ways; either using centralized or distributed registries.

Centralized Provision

The centralized service provision consists in scattering service registries in specific servers (i.e., registry servers) of the network. Therefore, for advertising a service, the provider must initially find which of these servers are available in the network. After that, it has to determine in which of them the service will be advertised (instead, the provider could advertise the service in all the available servers). When a client wants to discover services, it must also find the registry servers available in the network, and then discover the services advertised in them. It is important to notice that, once the registry servers are found, unless they become unreachable, clients and providers do not need to discover them anymore. Jini and Web Services are some of the current technologies that support centralized service provision. In Figure 3 we illustrate an example of such an approach. In such a figure, services are advertised by hosts A and B (Step 1), in a single registry server, host C (we are considering that the clients have already found the registry server). After that, each advertisement is stored in the service registry maintained by host C (Step 2). Also considering that host D has already found the registry server, it is then able to inquiry such server for the advertised services (Step 3). When a relevant service is found (Step 4), host D can interact with its provider, in this case host A, to retrieve the proxy for the service (Step 5).

A problem with this approach is that, if all these servers are off the network, services can not be discovered, even when their providers are available. In these cases a possible solution could be the election of new servers from the moment that is detected that the others are no longer available. Furthermore, the centralized service provision raises the same problem of the pull-based one concerned with the validity of the service advertisements. In this way, the same, or at least similar, solutions can be applied here.

As the centralized service provision requires the existence of registry servers, it is not well suited for highly dynamic

Figure 3. Example of a centralized service provision

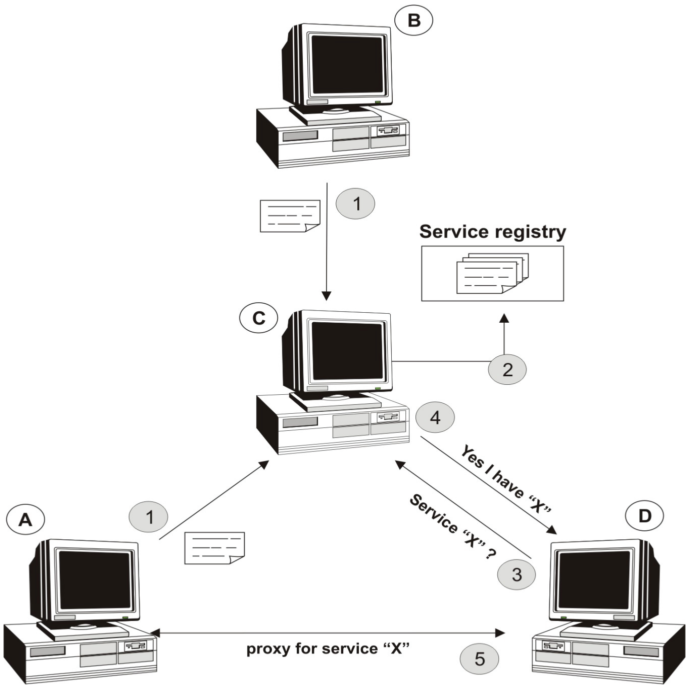
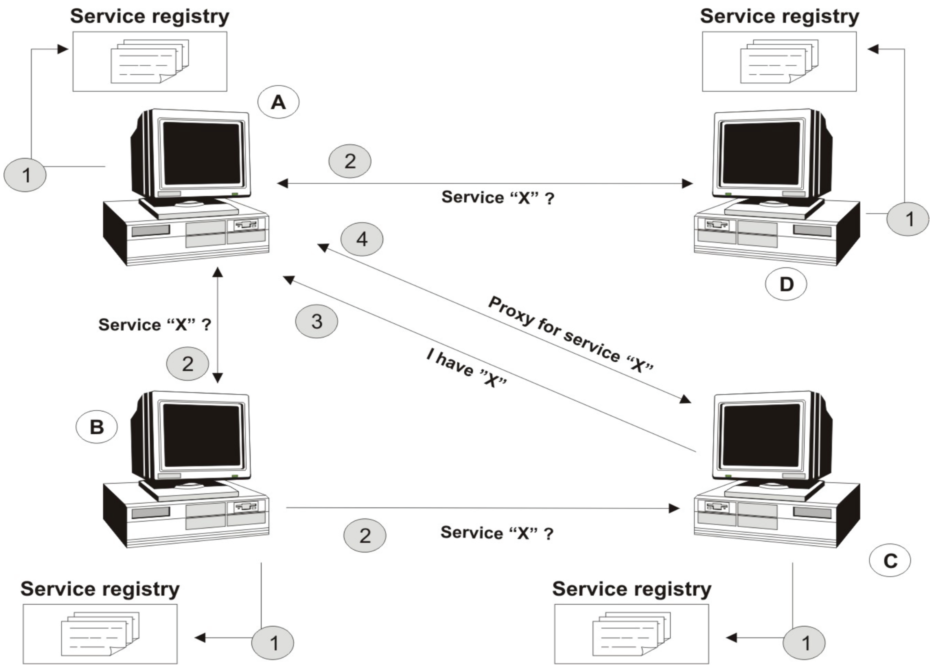


Figure 4. Example of a distributed service provision



pervasive environments. In these cases, as nodes join and leave the environment all the time, there would be too many changes in the current registry servers, which would in turn degrade the provisioning of services. On the other hand, this approach is very useful for environments equipped with wired network. In such environments, services can be deployed in the wired network and thus be accessed by mobile clients through wireless links. In environments populated with lots of mobile clients, this is certainly a good choice, as the bandwidth available in the wired network could support a great number of accesses.

Distributed Provision

In the distributed service provision, services are advertised in registries located in each host. Undoubtedly, in this approach the advertising task is easier to be performed than in the other ones, as it does not involve sending advertisements to central servers or directly to the other hosts. However, service discovery is more complicated, as it must be performed by inquiring each available host for the needed services. As no centralizer hosts are necessary for advertising services, discovery is possible whenever a client and a provider are present in the network. An example of the distributed service provision approach is illustrated in Figure 4. Initially, each host advertises its services (Step 1). Once a client needs to perform service discovery, in our example host A, it asks every host in the network for the needed service (Step 2). It is important to note the possibility of redirecting the service discovery request to hosts that are not reachable from the client. In Figure 4 this is performed by host B when it redirects the request of host A to C. When a host has a service matching the client's needs, it sends a notification (Step 3). From this notification, the client can then retrieve a proxy to the service (Step 4).

The major problem with this approach is associated with the protocols for performing service discovery. As any node in the network is a potential service provider, the discovery protocol must be well designed, in order to cover all hosts of the network. If any host is missing in the discovery process, it is possible that a relevant service may not be found. In *ad hoc* pervasive environments, one possible solution to this problem is first to discover the hosts in the vicinity, and then to retrieve the relevant services they provide. However this solution only performs well in small networks, where the available hosts can be discovered from any other one. Another point to be considered is about the network congestion that such service discovery protocols can cause. As the search should include all hosts in the network, the protocol must apply techniques for avoiding flooding the network. Obviously, this is not a serious problem in wired networks, but considering the wireless ones, it must be strictly taken into account. A good usage scenario of the distributed provision approach is a decentralized pervasive environment where the edge of the network is formed by mobile clients and its core is populated by service providers connected through a wired network.

SERVICE ORIENTED TECHNOLOGIES

In this section we present the main technologies related to the provision of services in pervasive computing environments.

Universal Plug and Play (UPnP)

The Universal Plug and Play (Richard, 2000) is an open architecture, which uses standard Internet protocols for pervasive peer-to-peer network connectivity (http://www.upnp.org). The UPnP protocol defines a set of steps, *addressing, discovery, description, control, eventing,* and *presenting*, which enables the automatic network configuration and service provision. Through the addressing, devices get their network address, which is performed by a dynamic host configuration protocol (DHCP). The discovery step consists of notifying the other hosts, through the push-based approach, about the services, and embedded devices, that a joining host provides. The discovery can also be performed in a distributed pull-based fashion. The next step, description, is about the description of a device, stored as an XML document. In such a document, it is kept, among other information, the description of the services that the device provides. Therefore, by discovering a device, it is possible to retrieve the services it provides, and then, through the control step, invoke their actions. Changes in a service can be notified to interested devices through the eventing step. Finally, through the presenting step, it is possible to load a specific URL, specified in the device description, in order to display a user interface for controlling a device.

Jini

Jini is a service-oriented Java technology based on a centralized pull-based approach (Waldo, 1999). Therefore, service advertisements are stored in central servers, which are named *lookup servers*. Jini uses the RMI (http://java.sun.com/products/jdk/rmi - Remote Method Invocation) protocol for all interactions involved in the advertisement, discovery, and invocation of services. When a client discovers and binds to a service, it is incorporated to the client by downloading the code of a proxy to the required service, named *remote control object*.

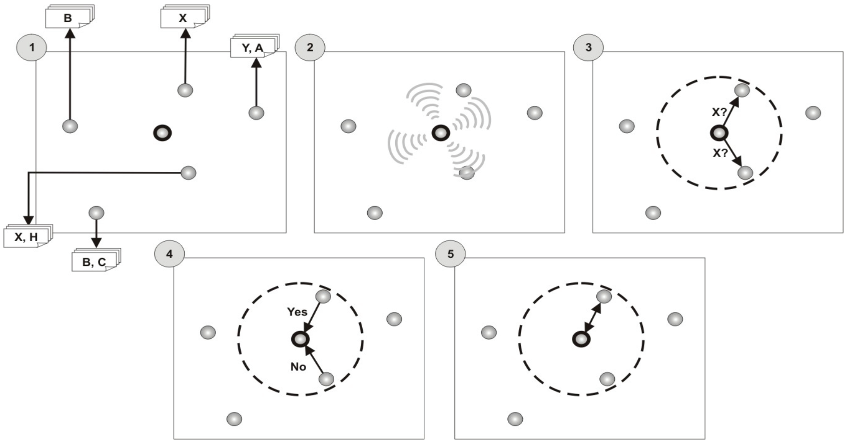
The Jini platform uses the concept of lease for controlling the access to the services. A lease is a sort of warrant that a client has for using a service during a specific period of time. When the lease expires the client needs to renew it with the provider if it wishes to continue using the service.

Bluetooth

Bluetooth is a standard for wireless communication among small devices within short distances (Johansson, Kazantzidis, Kapoor, & Gerla, 2001), defining higher-level protocols for both host and service discovery (http://www.bluetooth.org). The discovery of services in the Bluetooth standard is defined by the service discovery protocol (SDP), which enables to enumerate the devices in the vicinity and retrieve the services they provide.

Bluetooth uses a distributed pull-based approach for service advertising and discovery. To this end, each device maintains its own service discovery database (SDDB), which is a registry where its services are advertised. Therefore, a Bluetooth device performs service discovery by querying the SDDBs of the devices around. These advertising and discovery processes are illustrated in Figure 5. Notice that, initially, all devices advertise their services on their respective SDDBs (1). Next, a client searches for all the Bluetooth devices on the range of its wireless interface (2). For each device found, the client sends a query about the availability of an interested service (3). The devices answer these queries

Figure 5. Service discovery in Bluetooth



by informing whether they offer the needed service or not (4). Once localized a device providing the desired service, the client can connect directly to such device and finally use the service (5).

FUTURE RESEARCH TRENDS

Although the first service-oriented pervasive computing solutions have been developed, much work has to be done yet. For example, the matching of the user's needs and the services' functionalities should be enhanced to improve the interaction of the user with the pervasive application. Still, problems remain in the context of service continuity. Solutions to this problem would enable the user to use a service continuously, as he/she walks through different environments. This is important because, sometimes, the service a client was using is not available in the new environment, and thus, some mechanism should allow it to use a similar one without, or at least with a minimum, of interruption.

These, and possibly other problems related to the provision of services in pervasive environments, must certainly be completely solved so that we can enjoy the full potential of merging service-oriented and pervasive computing.

CONCLUSION

The service-oriented paradigm has proved to be an important element in pervasive computing systems, in order to provide anytime and anywhere access to services. Its dynamic binding feature enables to build applications powered with on-demand extensibility and adaptability, two important elements of any pervasive system.

Given this trend, in this chapter we have tried to present an overview of service provision in pervasive computing environments. More precisely, we have showed an introduction to the main characteristics, challenges, and solutions concerning the way that services are advertised, discovered, and used in pervasive environments. Although we presented concepts at an introductory level, we believe they may serve as a good source of knowledge, helping both students and researchers involved with these fields.

REFERENCES

Bellur, U., & Narendra, N. C. (2005). Towards service orientation in pervasive computing systems. In *International Conference on Information Technology: Coding and Computing* (Vol. II, pp. 289-295). Las Vegas, NV.

Costa, P., Coulson, G., Mascolo, C., Picco, G. P., & Zachariadis, S. (2005). The RUNES middleware: A reconfigurable component-based approach to networked embedded systems. In *Proceedings of the 16th IEEE International Symposium on Personal Indoor and Mobile Radio Communications*. Berlin, Germany: IEEE Communications Society.

Huhns, M. N., & Singh, M. P. (2005). Service oriented computing: Key concepts and principles. *IEEE Internet Computing*, *9*(1), 75-81.

Johansson, P., Kazantzidis, M., Kapoor, R., & Gerla, M. (2001). Bluetooth: An enabler for personal area networking. *IEEE Network, 15*(5), 28-37.

Loureiro, E., Bublitz, F., Oliveira, L., Perkusich, A., Almeida, H., & Ferreira, G. (2006). A flexible middleware for service provision over heterogeneous pervasive networks. In *Proceedings of the 4th International Workshop on Middleware for Mobile and Distributed Computing*. Niagara Falls, NY: IEEE Computer Society.

McGovern, J., Tyagi, S., Stevens, M., & Mathew, S. (2003). Service oriented architecture. In J. McGovern, S. Tyagi, M. Stevens, & S. Mathew (Eds.), *Java Web services architecture* (pp. 35-63). San Francisco: Morgan Kaufmann.

Nickull, D. (2005). *Service oriented architecture* (White Paper). San Jose, CA, USA: Adobe

Systems Incorporated. Retrieved March 28, 2006, from: http://www.adobe.com/enterprise/pdfs/Services_Oriented_Architecture_from_Adobe.pdf

Papazoglou, M. P. (2003). Service-oriented computing: Concepts, characteristics, and directions. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering* (pp. 3-12). Rome: IEEE Computer Society

Robinson, J., Wakeman, I., & Owen, T. (2004). Scooby: Middleware for service composition in pervasive computing. In *Proceedings of the 2nd Workshop on Middleware for Pervasive and Ad hoc Computing*. Toronto, Canada.

Richard, G.G. (2000). Service advertisement and discovery: Enabling universal devicecooperation. *IEEE Internet Computing*, *4*(5), 18-26.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal Communication, 8*(4), 10-17.

Sivaharan, T., Blair, G., & Coulson, G. (2005, October). GREEN: A configurable and reconfigurable publish-subscribe middleware for pervasive computing. In *Proceedings of the International Symposium on Distributed Objects and Applications* (Vol. 3760, pp. 732-749). Agia Napa, Cyprus: Springer Verlag.

Turban, E., Rainer, R. K., & Potter, R. (2005). Mobile, wireless, and pervasive computing. *Information Technology for Management: Transforming Organizations in the Digital Economy* (pp. 167-206). New York: John Wiley & Sons.

Waldo, J. (1999). The Jini architecture for network-centric computing. *Communications of the ACM*, *42*(7), 76-82.

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 94-104.

KEY TERMS

Pervasive Computing: The vision conceived by Mark Weiser which consists of world where computing will be embedded in every day objects.

Service: a software entity that can be integrated to a remote distributed application.

Service-Oriented Computing: The newest paradigm for distributed computing, where applications should be built by dynamically integrating services.

Service Advertisement: The element used for publishing and discovering a service.

Service Client: The one wishing to use a service.

Service Provider: The one that offers services.

Service Registry: The place where services are published.

Short Message Service (SMS) as an Advertising Medium

Shintaro Okazaki
Autonomous University of Madrid, Spain

INTRODUCTION

The proliferation of the Internet-enabled mobile device has extended into many parts of the world. Collectively, the mobile-network operators paid more than $100 billion for licenses to operate “third-generation” (3G) networks, which were among “the largest bet in business history on the introduction of a new technology” (Economist, 2005). This drastic move has been most illustrated by the use of short message service (SMS) and multimedia messaging service (MMS) by mobile users. For example, a recent survey indicates that SMS in the Asia-Pacific region will increase to up to 75% of mobile subscribers in 2006 (IDC Asia/Pacific, 2003). As a result, marketers and agencies are increasingly interested in taking advantage of this growth, by incorporating SMS advertising as part of an integrated marketing communications (IMC) strategy. However, there has been little academic research on mobile advertising, perhaps because its growth is still in an early stage and the technological infrastructure varies across markets. The study has two objectives: (1) to identify the factors influencing MNCs’ managerial intention to adopt SMS advertising, and (2) to test a statistical relationship between these factors and managerial intention to use SMS advertising. To this end, we conducted telephone interviews of senior executives of MNCs operating in European markets.

CONCEPTUAL FRAMEWORK AND HYPOTHESES

Branding Technique

In an environment where building the brand is a fundamental goal for many managers, the need to build brand equity is likely to be at the center of many marketing decisions. Firms using SMS-based campaigns can attract consumer attention and produce consumer responses to a much greater degree than via other direct marketing channels, because SMS has been claimed to be an effective tool in building and testing customer loyalty by developing demographic databases (Mylonopoulos & Doukidis, 2003). From an industry perspective, McDonald’s conducted a text-messaging campaign in conjunction with a popular TV song contest in the UK, offering concert tickets and backstage passes, while entry in the Coca-Cola Grand Sweepstakes Competition was offered to U.S. college students who sent a text message to a number printed on a Diet Coke can (Dano, 2002).

Facilitating Conditions

Lu, Yu, Liu, and Yao (2003) suggest that facilitating conditions are one of the most important determinants, along with the ease of using wireless Internet. In this light, the integration of competing standards and fragmented systems across countries, cross-network support for SMS, and higher connection speeds are all necessary conditions for a wider transmission of mobile advertising. In addition, the availability of Web-enabled mobile handsets with 2.5G or 3G functionality would significantly affect the adoption of MMS-based (multimedia message services) campaigns. In this light, a wider selection of handsets must be available, to enable consumers to choose their preferred combination of necessary functions and diverse features.

Location-Based Services

The satellite-based global positioning system (GPS) offers the ability to tailor services and promotional offers to individual consumers’ needs, by locating their position (Sadeh, 2002). Mobile handset makers and content providers are increasingly attracted by the commercial feasibility of applying GPS to their service. For example, on an extended menu of i-mode, “i-area” includes a diverse range of location-based services: weather news, restaurant guide, local hotel information, zoomable maps with an address finder function, and traffic updates and estimation of travel times. This facility would give MNCs strategic leverage in mobile marketing, because individuals’ behavior and receptiveness to advertising is likely to be influenced by their location and time, and marketers can thus induce impulse buying by providing the right information for the right place (Barnes, 2002).

Connection Costs

Another important factor is the concept of connection costs. For example, to send or receive one megabyte of data on 2.5G i-mode costs 32 euros (0.3 yen) per packet. At a rate

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

of 19 euro cents per 160-character SMS message, European consumers would have to pay 1,356.98 euros to send one megabyte of data by SMS, or approximately 62 times as much as the Japanese pay (Scuka, 2003). In addition, European mobile operators have passed on to consumers the additional costs incurred in obtaining 3G spectrum licenses, and this has made any dramatic price reduction impossible (Baldi & Thaung, 2002). Such cost factors adversely affect mobile players' revenues.

Public Regulation

The idea behind mobile *advertising* is very similar to e-mail on the wired Internet, but with one big difference: it is "opt-in." This function is essential to give users total control over what they receive, because consumers' demand for highly personalized messages has to be reconciled with their desire for privacy (Sadeh, 2002). The Mobile Marketing Association (MMA) has attempted to establish industry guidelines for mobile marketers, as follows: (1) MMA members should not send mobile advertising without confirmed opt-in, and (2) such opt-in subscriber permission is not transferable to third parties without explicit permission from the subscriber (Petty, 2003).

Lifestyle and Habits

In general, European consumers habitually commute by car, and this provides fewer incentives to access the mobile Internet (Baldi & Thaung, 2002). In addition, a systematic "word-of-mouth" helped the rapid diffusion of i-mode in Japan, especially given the "normative beliefs attributed to significant others (friends, colleagues, or family members) with respect to adopting or continuing to use the technology" (Barnes & Huff, 2003). This may partially explain a high subscription rate (almost 75%) to e-mail newsletters among i-mode users, and this makes acceptance of mobile advertising much easier. However, this factor is unlikely to be present in many European countries, which are characterized as more individualist than Asian countries.

On the basis of the preceding discussions, the following hypotheses were formulated to test the principal thesis of the research:

- **H1:** MNCs' intention to adopt SMS-based advertising is directly and positively associated with branding technique.
- **H2:** MNCs' intention to adopt SMS-based advertising is directly and positively associated with facilitating conditions.
- **H3:** MNCs' intention to adopt SMS-based advertising is directly and positively associated with location-based services.
- **H4:** MNCs' intention to adopt SMS-based advertising is directly and negatively associated with connection costs.
- **H5:** MNCs' intention to adopt SMS-based advertising is directly and negatively associated with public regulation.
- **H6:** MNCs' intention to adopt SMS-based advertising is directly and negatively associated with lifestyle and habits.

METHODOLOGY

Questionnaire Items and Measures

A structured questionnaire was prepared, drawing on prior literature. A majority of the items were originally developed for this study, because of the scarcity of empirical research on mobile advertising. Each item was measured on a Likert-type five-point scale. A five-point scale was preferred to a seven-point scale, because telephone interviews were used, rather than a mail or other form of paper-and-pencil survey. This method was considered more appropriate because mobile advertising is still in its infancy, and company executives may not be able to make fine distinctions regarding their attitudes on this topic. During the telephone interview, interviewers followed a script. However, respondents were free to ask questions whenever they encountered definitional problems.

Multinational Corporations

With regard to Japanese firms, the selection was based on the *Multinational Companies Database.* The database was created by the Research Institute for Economics and Business Administration at Kobe University (2003) and includes Japanese companies listed in the first section of the Tokyo Stock Exchange with foreign direct investment in more than five countries (Kobe University, 2003). American firms were chosen from The Forbes 500 (*Forbes*, 2003a). Finally, European firms were singled out from The Forbes International 500 (*Forbes*, 2003b), because this list indicates the nationality of each firm. Regardless of nationality, however, companies associated with aerospace and defense, food and drug retail chains, forestry and fishery, general public utilities, health care providers, heavy machines, industrial goods, local banking and insurance, metals and mining, and oil and gas extraction were excluded. Next, firms operating in Spain were identified. As a result, 43 Japanese, 47 American, and 31 European firms' Spanish subsidiaries were identified.

Table 1. Regression analysis and hypotheses testing

Hypotheses	Independent Variables		Standardized β		Results
H1	Branding technique		.553	**	Supported
H2	Facilitating conditions		.325	**	Supported
H3	Location-based services		.023		Rejected
H4	Connection costs		-.397	**	Supported
H5	Public regulation		.125		Rejected
H6	Lifestyle and habits		-.115		Rejected
		R^2	.598	**	
		ΔR^2	.013		
		ΔF	1.509		

Note: Dependent variable = MNCs' intention to use mobile advertising
** *Significant at $p < .001$*
* *Significant at $p < .05$*

Telephone Interview

Telephone interviewing was considered appropriate because of the novelty of the research subject. It was expected that interviewers would be able to clarify doubts or answer any questions that interviewees might have regarding mobile communications. To this end, four bilingual assistants were employed (two Spanish and two Japanese, all fluent in English). During the second and third weeks of February 2004, intensive training was provided so that the assistants could gain sufficient skills and knowledge to conduct the telephone interview. The actual interviewing was carried out during March 2004, under the supervision of the researcher. It was established that when the target executives were absent or unavailable for interview, assistants had to ask: (1) for an appointment for the next phone call, or (2) about the availability of the person next in seniority in the marketing department to the target executive. As a result, a total of 53 interviews was conducted, with 27, 16, and 10 respondents from Japanese, American, and European firms, respectively. The response rate was 43.8%.

FINDINGS

First, an exploratory factor analysis with Equamax rotation with Kaiser Normalization was carried out. The rotation, converged in 12 iterations, produced a clear-cut six-factor solution with a cut-off value of .50. Only factors with eigenvalue greater than 1 were retained. It should be noted that the proposed construct "connection costs" was merged into a mixed construct "security and costs." However, because of the exploratory nature of the study, it was considered acceptable to use this six-factor solution for the subsequent analysis. The extracted factors explain 68.6% of the total variance, and the level of loading is consistently high across the six factors. Factor scores were retained as variables with the Anderson-Rubin method to minimize the level of multicollinearity, for the use of regression analysis. The reliability was calculated with Chronbach's alpha for each construct. The scores range from .60 to .85, exceeding the cut-off point of .60 suggested by Hair, Anderson, Tatham, and Black (1998). Next, the hypotheses were tested by performing regression analysis with a step-wise method. Each of six independent variables (i.e., factor scores) was regressed on the dependent variable, "MNCs' intention to use mobile advertising," in order of their expected contributions. The results of regression analysis are shown in Table 1.

DISCUSSION

This study aims to identify MNCs' principal perceptions of SMS-based push-type mobile advertising and their intention to use it. On the basis of the data obtained from 53 MNCs, our principal propositions were tested by multiple regression analysis. The results were mixed: only half of the six hypotheses gained empirical support. The regression analysis identified branding technique, facilitating conditions, and connection costs as the three primary predictors influencing MNCs' intention to use mobile advertising. The contribution of branding technique in particular is substantial, indicating that MNCs are likely to perceive mobile advertising as an effective branding tool to increase brand awareness and image. Also, technological infrastructure and the availability of sophisticated mobile handsets are prerequisites for mobile marketing. As expected, unfavorable mobile Internet pricing negatively affects the MNCs' intention to use mobile advertising. On the other hand, the contributions of location-based services, public regulation, and lifestyle and habits are not only statistically insignificant, but also trivial in terms of the coefficient magnitude. One reason why location-based services were not identified as a significant factor is that the GPS system is not as widespread in Europe as it is in Japan. In addition, many Scandinavian firms, leaders of sophisticated mobile Internet service practitioners, were not included in the study. Admitting the danger of simple generalization, the findings of this study may imply that MNCs are concerned to a lesser extent with regulatory and cultural impediments to adopting mobile advertising.

REFERENCES

Baldi, S., & Thaung, H. P. P. (2002). The entertaining way to m-commerce: Japan's approach to the mobile Internet—A model for Europe? *Electronic Markets, 12*(1), 6-13.

Barnes, S. J. (2003). Wireless digital advertising: Nature and implications. *International Journal of Advertising, 21,* 399-420.

Barnes, S. J., & Huff, S. L. (2003). Rising sun: i-mode and the wireless Internet. *Communications of the ACM, 46*(11), 79-84.

Barwise, P., & Strong, C. (2002). Permission-based mobile marketing. *Journal of Interactive Marketing, 16*(1), 14-24.

Dano, M. (2002). Coke, Toyota, McDonald's test mobile advertising. *RCR Wireless News, 21*(46), 8.

Forbes. (2003a). *The Forbes 500s.* Retrieved November 3, 2003, from http://www.forbes.com/2003/03/26/500sland.html

Forbes. (2003b). *The Forbes International 500.* Retrieved November 15, 2003, from http://www.forbes.com/2003/07/07/internationaland.html

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis.* Upper Saddle River, NJ: Prentice Hall.

Kobe University. (2003). *Multinational Companies Database.* Available by permission of Research Institute for Economics and Business Administration of Kobe University. Retrieved January 11, 2004, from http://www.rieb.kobe-u.ac.jp/liaison/cdal/takokuseki/dbenterprises.html

Lu, F., Yu, C. S., Liu, C., & Yao, F. E. (2003). Technology acceptance model for wireless Internet. *Internet Research, 13*(3), 206-222.

Mylonopoulos, N. A., & Doukidis, G. I. (2003). Introduction to the special issue: Mobile business: Technological pluralism, social assimilation, and growth. *International Journal of Electronic Commerce, 8*(1), 5-22.

Petty, R. D. (2003). Wireless advertising messaging: Legal analysis directly and public policy issues. *Journal of Public Policy and Marketing, 22*(1), 71-82.

Robinson, J. P., Shaver, P. R., & Wrightsman, L.S. (1991). Criteria for scale selection and evaluation. In J. P Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 1-16). San Diego: Academic Press.

Sadeh, N. (2002). *M-commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

Scuka, D. (2003). *How Europe really differs from Japan.* Retrieved February 11, 2004, from http://www.mobiliser.org/article?id=68

KEY TERMS

Barcode Mobile Coupon: Mobile barcoding can be used in the form of a picture SMS which is delivered to a mobile phone. Recipients save the image, arrive at the destination, and present their barcode SMS to be scanned.

i-mode: A broad range of Internet services for a monthly fee of approximately three euro, including e-mail, transaction services (e.g., banking, trading, shopping, ticket reservations, etc.), infotainment services (e.g., news, weather, sports, games, music download, karaoke, etc.), and directory services (e.g., telephone directory, restaurant guide, city information, etc.), which offers more than 3,000 official sites accessible through the i-mode menu.

Push Messaging Service: Various forms of messaging services are generally offered in mobile Internet. For example, SMS and WAP Push messaging generally allow users to send 100-160 characters, while mobile e-mail in Japanese i-mode allows up to 1,000 characters.

SPAM: Unsolicited or undesired bulk electronic messages. Because of the development of anti-SPAM programs, they are often deleted without being opened.

SPIM: A variation of SPAM through instant messaging systems.

Shot Boundary Detection Techniques for Video Sequences

H. Koumaras
N.C.S.R., Demokritos, Greece

G. Xilouris
N.C.S.R., Demokritos, Greece

E. Pallis
Technological Educational Institute of Crete, Greece

G. Gardikis
University of the Aegean, Greece

A. Kourtis
N.C.S.R., Demokritos, Greece

INTRODUCTION

The advances in digital video encoding and compression techniques that achieve high compression ratios by exploiting both spatial and temporal redundancy in video sequences have made possible the storage, transmission, and provision of very high-volume video data over communication networks.

Today, a typical end user of a multimedia system is usually overwhelmed with video collections, facing the problem of organizing them in a browsing-friendly way. Thus, in order to allow an efficient exploitation and browsing of these video-anthologies, it is necessary to design techniques and methods for content-based search and access. Therefore, the issue of analyzing and categorizing the video content by retrieving highly representative optical information has been raised in the research community.

Thus, the current trend has led to the development of sophisticated technologies for representing, indexing, and retrieving multimedia data. A common first step towards this is the segmentation of a video sequence into elementary shots, each comprising a sequence of consecutive frames that record a video event or scene continuously in the spatial and temporal domain. Moreover, these elementary shots appear as they have been captured by a single camera action. Two adjacent elementary streams are divided by a *shot boundary* or *shot transition,* also known as scene cut, when the change of video content occurs over a single frame, or *gradual shot boundary,* when the changes occur gradually over a short sequence of frames (e.g., dissolve, fade in/out, etc.) (Lu & Tan, 2005).

In general, gradual transitions are more demanding in detection than abrupt scene cuts, because they must be distinguished from regular camera operations that cause similar temporal variances and usually trigger false detections. Especially for video content with high spatial and temporal activity level, the detection of gradual scene changes becomes even more challenging (Hampapur, Jain, & Weymouth, 1995).

Hence, the goal of this temporal video segmentation is to divide the video stream into a set of meaningful and manageable segments that are used as basic elements for indexing. Further analysis may be performed, such as representation of the video content and event identification.

In future multimedia systems, the offered video services will be provided in the form of MPEG-21 digital items, which integrate a typical encoded media clip along with its XML-based metadata descriptors, enabling in this way advanced search and retrieve abilities. Also future multimedia implementations will adapt MPEG-21 schema, which means that upcoming media recorders must be able to automatically create video content indexing.

This chapter will outline the various existing methods of boundary shot and scene change detection.

BACKGROUND

A primitive typical approach to indexing video data was the manual creation of textual annotations along with time headers in the metadata of a media file. However, such a human-based method is time consuming and practically not applicable. Moreover, such methods suffer from the subjectivity of the human operator during the textual description.

Therefore, it is necessary to develop an integrated framework for automatic extraction of the most character-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

istic frames of a video sequence, which will finally enable the efficient indexing and description of a video sequence. More specifically, by developing methods that enable the automatic build of a scene-access menu for a video clip, the viewer may use this index for quick access at a specific scene or for performing scene searches.

Several approaches have been proposed in the literature for automatic video indexing, which can be basically categorized as methods for temporal segmentation in an uncompressed or compressed video domain (Koprinska & Carrato, 2001; Lienhart, 1999; Dailianas, Allen, & England, 1995).

Thus, the various temporal video segmentation methods for each class (i.e., uncompressed/compressed) will be discussed in the following sections.

SHOT BOUNDARY DETECTION IN UNCOMPRESSED DOMAIN

Video segmentation in an uncompressed domain includes all the boundary shot detection methods that perform using metrics and mathematical models on the uncompressed/spatial video signal. Most existing methods detect shot boundaries of video based on some change of the video content on the visual domain between consecutive frame pairs. If the measured change is above a predetermined threshold, then a shot boundary is assumed and reported.

Based on the metrics nature that is used to detect the differences between successive frames, the algorithms can be generally classified into the following classes: pixel-based, block-based, and histogram-based (Zhang, Low, Gong, & Smoliar, 1994, 1995).

Pixel-Based Methods

Pixel-based methods evaluate the differences in luminance or color domain between pixel values of successive frames (Kikukawa & Kawafuchi, 1992). Hence, a per pixel comparison is performed between frame pairs. Depending on the measured difference from the pixel-based comparison, a scene cut is detected and reported if the calculated difference is above a pre-defined threshold value. Otherwise no scene change is reported. The sensitivity and the efficiency of the pixel-based methods are strongly related to the selection of the reference threshold.

Block-Based Methods

In contrast to the aforementioned pixel-based methods, where the whole frame of a video movie is taken under consideration for the scene change detection and the corresponding measured difference in the pixel values, either in color or luminance domain, in block-based methods each frame is divided into blocks that in turn are compared to their corresponding blocks in the successive frame (Kasturi & Jain, 1991; Shahraray, 1995). More specifically, in contrast to the aforementioned pixel-based techniques, where the critical unit is the number of pixels whose difference is above a threshold value, these methods report a scene change, when the number of changed blocks is greater than a predefined threshold.

Histograms Comparisons

The aforementioned categories exploit pixel comparison in order to derive a decision. On the contrary, histogram-based methods exploit the fact that a set of frames that belong in the same scene retain, in general unchanged, their luminance- or color-level histograms. A luminance- or color-level histogram of a frame depicts the density of the number of pixels that have specific luminance or color value.

As has been described, the majority of the aforementioned methods are implemented based on metrics of the uncompressed video domain, utilizing a common framework: a similarity measurement between successive frames.

SHOT BOUNDARY DETECTION IN COMPRESSED DOMAIN

Multimedia applications that distribute audiovisual content over communication networks (such as video-on-demand (VOD) and real-time entertainment streaming services) are based on digital encoding techniques (e.g., MPEG-1/2/4 and H.261/2/3 standards) that achieve high compression ratios by exploiting the spatial and temporal redundancy in video sequences. Most of the standards are based on motion estimation and compensation, using the block-based discrete cosine transformation (DCT). The use of transformation facilitates the exploitation in the compression technique of the various psychovisual redundancies by transforming the picture to a domain where different frequency ranges with dissimilar sensitivities at the human visual system (HVS) can be accessed independently.

The DCT operates on a X block of N X N image samples or residual values after prediction and creates Y, an N X N block of coefficients. The action of the DCT can be described in terms of a transform matrix A. The forward DCT is given by:

$$Y=AXA^T$$

where X is a matrix of samples, Y is a matrix of coefficients, and A is an N X N transform matrix. The elements of A are:

$$A_{ij} = C_i \cos\frac{(2j+1)i\pi}{2N},$$

where

$$C_i = \begin{cases} \sqrt{1/N}, i = 0 \\ \sqrt{2/N}, i > 0 \end{cases} \qquad (1)$$

Therefore the DCT can be written as:

$$Y_{xy} = C_x C_y \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} X_{ij} \cos\frac{(2j+1)y\pi}{2N} \cos\frac{(2i+1)x\pi}{2N} \qquad (2)$$

Afterwards in the encoding chain, quantization of the aforementioned DCT coefficients is performed, which is the main reason for the quality degradation and the appearance of artifacts, like the ‘blockiness’ effect.

Several methods for shot boundary detection in the compressed domain have been developed. According to Koprinska and Carrato (2001), they can be classified into the following categories, depending on the used metric:

- **DCT Coefficients:** The temporal video segmentation methods based on DCT coefficients apply a comparison technique to the DCT coefficients of the corresponding successive video frames. The difference metric is somewhat similar to the aforementioned pixel-based metric, where a scene change is detected and reported when the measured difference exceeds a specific threshold value (Zhang et al., 1994, 1995).
  It must be noted that these methods can be applied only on intra-coded frames of a DCT-based coded signal, because only they are fully encoded with DCT coefficients. Thus, the processing requirements may be low, but the temporal accuracy of the detected frame drops dramatically, and it is highly dependent on the intra-frame periodicity.
- **DC Terms:** The DC term is a scaled version of the average value for each block and thus the DC terms are directly related to the pixel domain. So, in a similar way to the uncompressed domain methods, the DC terms-based metrics measure the DC terms differences between successive frames. Again a frame is reported as shot boundary, if the aforementioned measurement is higher than a pre-defined threshold (Yeo & Liu, 1995).
- **DC Terms, Macroblock (MB) Coding Mode:** This is a hybrid method in which, except from the aforementioned DC terms, the type of the macroblock (MB) coding is taken under consideration as well. When a scene change takes place, then some macroblocks of an inter-coded frame may be intra-coded due to limited reference options, demonstrating where scene change occurs (Meng, Juan, & Chang, 1995).
- **MB Coding Mode and Motion Vectors (MVs):** Similarly to the previously described method, a hybrid model is exploited, where it takes under consideration both the MB coding mode and the MV information of the encoded video sequence.
- **MB Coding Mode and Bit Rate Information:** Finally this method uses both bit rate information and motion-predicted MB types in order to derive accurate estimation of the scene changes. The forced intra-coding of some MBs over a scene change increases the deduced bit rate due to the inefficiency of the intra-coding method.

Similarly to the shot change detection methods of the uncompressed domain, this section has shown that also the methods of the compressed domain exploit analogous frameworks at their implementation, which is based on the comparison of the calculated metric between successive frame pairs.

FUTURE TRENDS

All the aforementioned methods use a threshold parameter in order to distinguish shot boundaries and changes. Thus, a common problem in shot boundary detection lies in the selection of an appropriate threshold for identifying whether a change is sufficiently large to signify a shot boundary or not (Lu & Tan, 2005). If a global threshold is used for the detection of shot boundaries over the whole video, then successful detection rate may vary up to 20% even for the same video content (O’Toole, Smeaton, Murphy, & Marlow, 1999). To improve the efficiency and eliminate this performance variation, some later works propose to use an adaptive threshold which can be dynamically determined based on video content (Lienhart, 1999; Dailianas et al., 1995). But even these methods require a lot of computational power in order to estimate successfully the appropriate threshold parameter, making their implementation a challenging issue, especially for real-time applications.

Thus, the research community faces the challenge of developing new techniques and methods for detecting scene changes over a video signal by eliminating the necessity of threshold parameters in the decision process.

Moreover, in order to allow a more efficient exploitation and browsing of video-anthologies, it is necessary to integrate these boundary shot detection techniques within content-based search and access methods, where the categorization of the video content is occurred by retrieving highly representative optical and semantic information. In this respect, the combination of frame extraction techniques and semantics will help towards the evolution of the current Web to the Semantic Web, where the browsing and searching of information will be based on semantic information.

CONCLUSION

This article outlines the various methods for detecting and extracting the scene changes from a video sequence. Depending on the metric that is exploited for the detection procedure, the methods that have been proposed are classified into two broad categories: those based on the uncompressed domain and those that exploit the metric of the compressed domain. Both the categories share the common drawback that they use threshold values for their decisions. Thus, the research community faces the challenge to develop new techniques that eliminate the use of threshold values, eliminating in this way the complexity and the computational requirements of the proposed methods.

ACKNOWLEDGMENTS

This article is carried out within the "PYTHAGORAS II" research framework, jointly funded by the European Union and the Hellenic Ministry of Education.

REFERENCES

Dailianas, A., Allen, R. B., & England, P. (1995). Comparison of automatic video segmentation algorithms. *Proceedings of SPIE* (Vol. 2615, pp.2-16).

Hampapur, A., Jain, R., & Weymouth, T. E. (1995). Production model based digital video segmentation. *Multimedia Tools and Applications, 1*(1), 9-46.

Kasturi, R., & Jain, R. (1991). *Dynamic vision* (pp. 469-480). IEEE Computer Society Press.

Kikukawa, T., & Kawafuchi, S. (1992). Development of an automatic summary editing system for the audio visual resources. *Transactions on Electronics and Information, 75-A*(2), 204-212.

Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication, 16,* 477-500.

Lienhart, R. (1999). Comparison of automatic shot boundary detection algorithms. *Proceedings of SPIE* (Vol. 3656, pp. 290-301).

Lu, H., & Tan, Y-P. (2005). An effective post-refinement method for shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(11), 1407-1421.

Meng, J., Juan, Y., & Chang, S.-F. (1995). Scene change detection in a MPEG compressed video sequence. *Proceedings of the SPIE International Symposium on Electronic Imaging* (Vol. 2417, pp. 14-25). San Jose, CA.

O'Toole, C., Smeaton, A., Murphy, N., & Marlow, S. (1999). Evaluation of automatic shot boundary detection on a large video suite. *Proceedings of the 2nd UK Conference on Image Retrieval: The Challenge of Image Retrieval* (pp. 1-13). Newcastle, UK.

Shahraray, B. (1995). Scene change detection and content-based sampling of video sequences. *Proceedings of SPIE* (pp. 2-13).

Tam, W. J., Stelmach, L., Wang, L., Lauzon, D., & Gray, P. (1995, February 6-8). Visual masking at video scene cuts. *Proceedings of SPIE, 2411*, (pp. 111-119). San Jose, CA.

Yeo, B., & Liu, B. (1995). Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology, 5*(6), 533-544.

Zhang, H. J., Low, C. Y., Gong, Y. H., & Smoliar, S. W. (1994). Video parsing using compressed data. *Proceedings of the SPIE Conference of Image and Video Processing II* (pp.142-149).

Zhang, H. J., Low, C. Y., Gong, Y. H., & Smoliar, S. W. (1995). Video parsing and browsing using compressed data. *Multimedia Tools and Applications, 1,* 89-111.

KEY TERMS

Bit Rate: A data rate expressed in bits per second. In video encoding the bit rate can be *constant,* which means that it retains a specific value for the whole encoding process, or *variable,* which means that it fluctuates around a specific value according to the content of the video signal.

Frame: One of the many still images which as a sequence composes a video signal.

Histogram: A luminance- or color-level histogram of a frame depicts the density of the number of pixels that have specific luminance or color value.

Multimedia: The several different media types (e.g., text, audio, graphics, animation, video).

Pixel: Considered the smallest sample of a digital image or video.

Shot: An unbroken sequence of frames taken continuously from a single camera.

Video Codec: The device or software that enables the compression/decompression of digital video.

Video Coding: The process of compressing and decompressing a raw digital video sequence.

Smartphone Acceptance among Sales Drivers

Jengchung V. Chen
National Cheng Kung University, Taiwan

INTRODUCTION

The objective of this research is to find out the acceptance of sales drivers in logistic industry to use the smartphone in their work. This research uses two methods to collect data: survey and experiment. This research integrates technology acceptance model (TAM) (Davis, 1989), self-efficacy (Bandura, 1982, 1986), and innovation diffusion theory (IDT) (Rogers, 1962) into the research model to find out the factors of sales drivers in logistic industry accepting the smartphone. This experiment is focused on three user groups: the employees of a business that implements smartphone usage, the employees of businesses that do not use smartphone, and the students that are currently studying at a department of transportation and communication management. The results will help us understand whether this technology should be integrated into the traditional logistics system, and get to know the pros and cons of this idea.

BACKGROUND

Freight businesses in the logistics industry perhaps have few examples of utilizing mobile devices thoroughly in their daily operations. Sales drivers who are responsible for distributing goods on time need to not only interactively exchange information with the headquarters, but also need to use spare time visiting their customers. Because of the needs to better serve their customers and other operational purposes, logistics businesses have their employees equipped with all sorts of devices like hand held terminal (HHT), bar code reader, GPS, and on board unit (OBU) to keep track of the goods. In addition, those drivers who for a long time have been considered as shippers now have another important role—salesperson.

SMARTPHONES

Since the invention of the telephone in 1876, peoples' lifestyles have been changed drastically as time passes. Then, Martin Cooper introduced a whole new level of communication by using the concept of wireless technology, called the cellular phone.

Since then, the cellular phone has been part of many people's lives, and nowadays almost everyone owns a cellular phone. The cellular technology has evolved so drastically that phones with high-resolution digital cameras, voice recording and digital assistant are very familiar to most people. And to satisfy business users, powerful handheld devices that run the smartphone operating system have been developed. All these gadgets are here so that they can make peoples lives easier. Smartphones support many features that are really helpful in the business sector, for example, this cellular can be associated with OBU (on board unit) to create a more efficient delivery system for logistics companies. This technology has greatly benefited logistics companies since the implication of this technology took place. And more and more functions are being included for example portable bar code scanners and real-time communications systems that can update detailed information amongst the driver and the office at all times. Smartphones not only provide useful business oriented functions; they also provide useful functions such as calendars, task planning and even high-speed mobile Internet over the 3G network (Valletti & Cave, 2002; Ralph, 2002; Funk, 1998).

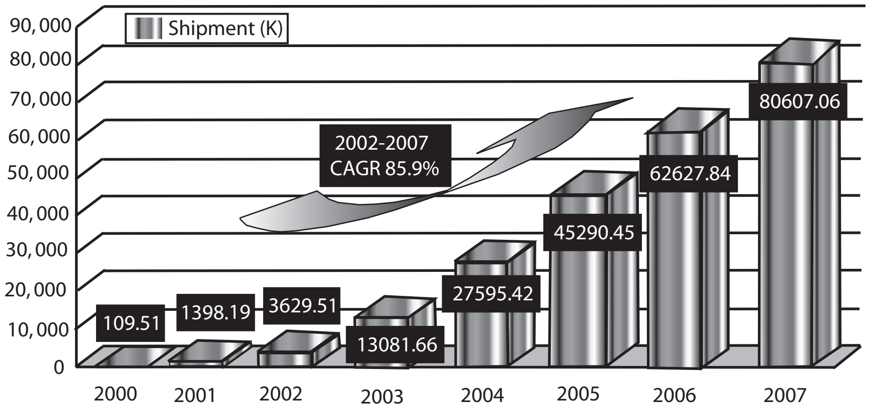
Different Categories of the Cellular Phones

Cellular phones are separated into three categories; the categories are sorted by the limitations of their functions shown as follows:

- **PDA Phones:** These kinds of phones support all the functions that a PDA can do, it is actually a whole PDA integrated into the phone. PalmOS, Symbian and Windows CE are examples of operating systems that are used in PDA phones. More and more software developers are developing operating systems that are becoming more powerful (e.g., Linux). Most of these PDA phones are even able to read and edit Word, Excel and even PowerPoint files, which is really convenient for business users who don't like to carry their computers around.
- **PIM Phones:** Known as the Personal Information Manager, it could also support features such as Outlook synchronization with a personal computer, but the functions are more limited compared to the PDA phones. PIM phones use a closed operating system; they do not support as many applications like the PDA phones.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. 2000-2007 shipment of smartphones (IDC)



- **Cellular Phones:** This is the most basic form of wireless phones, which usually offer simple address book functions, messaging, GPRS, WAP, MMS, and video calling. Some of these phones might support Java applets, but the functions are still very limited.

Operating System

There are currently many operating systems for mobile phones on the market, namely Linux, Windows Mobile, Symbian and Palm. Microsoft Windows Mobile and Symbian are most commonly used; this is because of the ease of use and the high compatibility of applications. Some of the operating systems' source codes for mobile phones are open for software developers to use, this can allow more software applications to be developed and allow a higher usage of mobile Smartphones (Gruber & Verboven, 2000; Harrison & Holley, 2001).

Differences between the different mobile phone operating systems are shown as follows:

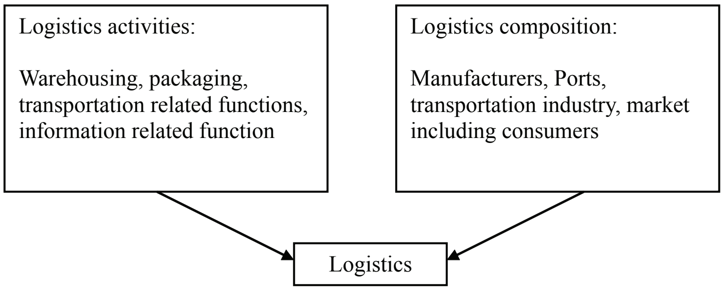
- Windows Mobile is separated into three different categories, namely Pocket PC, Pocket PC Phone edition and Smartphone; these are all developed by Microsoft. All of the three operating systems are very powerful and can support vast amounts of applications just like the PDA and even some PC applications.
- Symbian operating system is very commonly used because of the ease of use of the system. It also has integration of other software-developing companies that are continuously developing new applications and thinking out new ideas to improve it.
- Linux is well-known for being free; this makes no exception for Linux on mobile phones. This operating system's source code is open to the public; it allows any kind of alteration to the system itself and allows any developer to develop applications, all for free. Linux also turns out to be one of the most stable operating systems on the mobile phone market and the PC market too. But the problem with this operating system is the low support for applications.
- Palm OS has very high usage in the PDA market; they have successfully integrated the PDA technology with mobile phone technology. They are famous for their highly efficient input method by using the touch screen, and the high support of applications. Most applications that can be used on a Palm OS PDA can be used on a mobile phone running Palm OS for mobile phones.

Because of the rapid growth of smartphones' technology, more and more people are switching from a PIM or a normal cellular phone to using a smartphone, because they realize that it really can make a difference in their busy life. Smartphone-based technology has also been integrated into many logistic systems. From the data gathered from the IDC (International Data Collection), it shows that the number of Smartphones shipped is growing annually, and it also has been forecasted that this trend will continue to grow.

LOGISTICS

In the old days the word logistics had to do with the military's operations, it mainly dealt with procurement, distribution, maintenance and replacement of material and personnel. Nowadays it mainly has to do with the flow of material from one place to another, used mainly in the transportation industry. Logistics operation can include many other functions such as warehousing, packaging and other information based functions.

Figure 2. The composition of logistics


Logistics activities:
Warehousing, packaging, transportation related functions, information related function
Logistics composition:
Manufacturers, Ports, transportation industry, market including consumers
Logistics

Logistics can be separated into four main categories:

- **Manufacturing Logistics:** Logistics activities starting from the manufacturer's location to the location of the market.
- **Sales Logistics:** Logistics activities from the market to the consumer.
- **Procurement Activities:** Logistics activities based on purchasing of raw materials or other products.
- **Recycling Logistics:** Logistics activities based on returned merchandise.

Examples of motor vehicle logistics:

- Door to door service
- Post office service
- 3rd party logistics service
- Specialized transportation service
- Transportation of dangerous or abnormal goods
- Transportation rental service

As seen from the above, motor vehicles still play an important role in logistics, motor vehicles can access almost any hard to reach area, and it does not need any kind of facility (e.g., railroad, airport, port etc.) to enable the usage. It also can carry a tremendous amount of cargo.

SMARTPHONE AND LOGISTICS

To achieve maximum efficiency in logistics, it is recommended that smartphone technology be integrated. The use of modern technology can be really helpful in logistics service: bar code usage, GPS tracking and GPS navigation facilities have been of great help to the logistics industry in the recent years. These technologies allow customers to use logistics services with more confidence, because they know every detail that they need to know about the parcel. This is achieved by combining bar code technology with real-time GPS tracking. Drivers no longer need to keep calling the logistics company to track your parcel status. GPS also can navigate the driver to the shortest route available with real-time data; this can save a lot of time because drivers are not always familiar with the area that they are at.

SETTINGS FOR THE EXPERIMENT

This experiment is based on statistical analysis of data that is gathered through a questionnaire. The questionnaire is based on data from logistics companies that have already integrated mobile smartphone technology into their systems. The study is to investigate a case of a logistics company in the freight business; it is one of the major logistics companies in Taiwan that provides overnight delivery service of parcels and mail. They have already integrated smartphone and GPS technology into their delivery system. Bar codes are scanned at every point in the delivery process, and the data scanned is sent through the GSM network to the main control center; therefore a customer's area allows one to track the status of the objects that are sent. GPS is used to navigate the driver, and to avoid delays caused by heavy traffic zones. GPS can also get the driver to the right location without getting lost.

ANALYSIS OF RESULTS

The study collected 30 samples for this analysis:

- 10 employees from logistics companies that use smartphone
- 9 employees from logistics companies that do not use smartphone
- 11 students from a transportation and communications management major

The entire groups' ages range between 19 and 50 with an average of 31. Level of education is from high school or higher.

T-Type Analysis

Statistic t-test is used to analyze the data collected and it is found that only two hypotheses had positive results. After the experiment of function comparison, the VOIP function is more important compared to GPS and barcode reading; barcode has been shown to be very important too.

Five hypotheses are proposed, and only one is rejected. The only one that is rejected is the GPS function; they do not think that this is a very important function. But the result shows us that there is a very high level of acceptance of students, and statistics shows that logistics companies might be implementing more modern technology in the future.

FUTUTE TRENDS

The demand for smartphone technology will keep increasing in the future as the technology matures. The acceptance of smartphones in the transportation industry is growing, mainly because it can greatly increase the efficiency of delivery and customer relationships. Later models of smartphones bind many functions together so that you don't need to carry a lot of different equipment for different functions; for example handhelds units have GPS, GPRS and some even have bar code reading integrated in just one handheld PDA device. And the prices of these products are dropping gradually making it more affordable and increasing the will for a company to use these products.

CONCLUSION

From the results of the experiment, it shows that:

1. VOIP function shows more acceptance than other functions such as calendar functions.
2. GPS function is more important than the presentation function; navigation can allow drivers to avoid a lot of problems on the road.
3. VOIP is better than the messaging function; voice communication is direct and instantly responsive unlike messages.
4. GPS is more important than the barcode reading function. The barcode reading system has to be integrated with other technology to be able to function, which could sometimes be inconvenient. GPS shows more importance.

It is also identified the different levels of importance between different user groups:

1. GPS is considered useful for logistics companies; with this technology anyone is able to deliver products even if they are unfamiliar with the area.
2. Barcode is very important to companies that are currently using it. A majority of companies that are not using it yet show that they are willing to try it out.
3. VOIP has shown to be important to everyone, because this function can allow companies to save a tremendous amount of money on phone bills.

The utilizing of smartphones can benefit all parties by an increase the efficiency, which is one of the most important factors in the transportation industry. As smartphone technology grows, users outside the industry can also make use of smartphone devices in their daily lives, for example, GPS can be installed in all kinds of vehicles even on some bicycles. One reason for the rapid growth of technology is the high demand for new ideas, there is new technology designed for everyone, and making good use of it can improve our quality of life.

More academic theories should be surveyed to investigate the sales drivers' intention to use such the new technology. Take TAM for example, it excludes Hartwick and Barki's (1994) findings, which says subjective regulation impacts to the new technology acceptance. Other researchers proposed many more factors to be included in the TAM model. Future studies should also take these factors in to account: social influence processes and cognitive instrumental processes (Venkatesh & Davis, 2000), trust (Gefen, 2000), task-technology fit model (Dishaw & Strong, 1999), perceived characteristics of innovation (Plouffe et al., 2001), social influence and behavior control (Taylor & Todd, 1995).

REFERENCES

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*, 122-147.

Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice Hall.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

Dishaw, M. T., & Strong, D. M. (1999). Extending the technology acceptance model with task-technology fit constructs. *Information & Management, 36*(1), 9-21.

Funk, J. L. (1998). Competition between regional standards and the success and failure of firms in the world-wide mobile communication market. *Telecommunication Policy*, *22*(4), 419-441.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega: The International Journal of Management Science, 28*(6), 725-737.

Gruber, H., & Verboven, F. (2000). The diffusion of mobile telecommunications service in the European Union. *European Economic Review*, *45*, 557-558.

Hartwick, J., & Barki, H. (1994). Explaining the role of user participation in information system use. *Management Science, 40*(4), 440-465.

Harrison, F., & Holley, K. A. (2001). The development of mobile is critically development on standards. *BT Technology Journal*, *19*(1), 32-37.

Plouffe, C. R., Hulland, J. S., & Vandenbosch, M. (2001). Research report: Richness versus parsimony in modeling technology adoption decisions - understanding merchant adoption of a smart card-based payment system. *Information Systems Research, 12*(2), 208-222.

Ralph, D. (2002). 3G and beyond—The applications generation. *BT Technology Journal*, *20*(1), 22-28.

Rogers, E.M. (1962). *Diffusion of innovations*. NY: Free Press.

Taylor, S., & Todd, P. (1995). Assessing IT usage the role of prior experience. *MIS Quarterly, 19*(4), 561-570.

Valletti, T. M., & Cave, M. (2002). Competitive in UK mobile communications. *Telecommunication Policy*, *22*(2), 109-131.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

KEY TERMS

Barcode: A series of vertical bars of varying widths, in which each of the digits zero through nine are represented by a different pattern of bars that can be read by a laser scanner.

Global Positioning System (GPS): A system of satellites, computers, and receivers that is able to determine the latitude and longitude of a receiver on Earth by calculating the time difference for signals from different satellites to reach the receiver.

Logistics: Operations that deal with the procurement, distribution, maintenance, and replacement of material and personnel.

On Board Unit (OBU): Portable electronic device similar to a PDA.

Operating System: Software designed to control the hardware of a specific data-processing system in order to allow users and application programs to make use of it.

Smartphones: Mobile phone that had PDA functions integrated into it.

3G: Wireless technology that provides high-speed data transfer and portable video phone call service.

Voiceover Internet Protocol (VoIP): Voice communication that is done through Internet: usually can reduce costs of international calling.

SMS-Based Mobile Learning

Krassie Petrova
Auckland University of Technology, New Zealand

INTRODUCTION

Students today combine study and work and expect significant cost and time savings from the use of information and communication technologies, including mobile communication. A strong interest in implementing mobile technologies in learning has emerged. Experiments with one of the most popular technologies—text messaging—have been reported in the literature (e.g., Stone & Briggs, 2002; Finn, 2004) with some including the development of blended learning models (Stone, Briggs, & Smith, 2002).

An early definition of mobile learning (m-learning) as "learning through mobile computational devices" can be found in Quinn (2000). Later, frameworks and research models were developed, providing guidelines for implementing suitable pedagogical approaches and for building services and applications relevant to a variety of mobile platforms and contextual settings (Garner, Francis, & Wales, 2002; Seng & Lin, 2004; Berth; 2005; Brown, 2005).

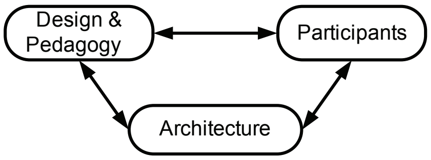
SMS (short message service, or text messaging) is an extremely popular and still growing 2G mobile data service, especially with young adults (Finn, 2004; Prensky, 2005; MMA, 2006; Grinter & Eldridge, 2003), which makes it suitable as a learning technology. This short article presents and illustrates the concepts of SMS-enabled m-learning, describing a series of SMS learning scenarios derived from the literature. The defining features of the scenarios are identified and discussed, including future trends.

BACKGROUND

Mobile learning is often referred to as a type of e-learning (Vavoula & Sharples, 2002; Leung & Chan, 2003; Seng & Lin, 2004). For the purposes of this article mobile learning is defined as a form of e-learning, which can take place anytime and anywhere through the use of a wireless and mobile communication device and the related network technology (Brown, 2005; Kukulska-Hulme, Evans, & Traxler, 2005; Wagner, 2005; Petrova, 2007).

An SMS scenario can be defined as a self-contained learning experience focused on a group of participants who act in a specific context and perform specific tasks to achieve knowledge acquisition oriented goals using the SMS mobile technology (Petrova & Sutedjo, 2004; Evans & Taylor, 2004). As text messaging is enabled on all types of 2G and 3G mobile phones, an SMS-learning scenario will be accessible to virtually any mobile phone user.

Figure 1. An SMS-based learning scenario framework



The framework in Figure 1 captures the main aspects from which researchers have described and evaluated m-learning, including SMS-based learning (Trifonova, 2003; Attewell, 2005; Riordan & Traxler, 2005; Silander & Rytkonen, 2005; Chinnery, 2006).

Architecture

Architecture deals with the specific mobile platform or platforms developed and used within a mobile learning scenario. The basic architecture would include access to the SMS provider network, an SMS-enabled cell phone and an SMS server or gateway (Petrova, 2007; Capuano, Gaeta, Miranda, & Pappacena, 2004). It might also include a number of auxiliary servers, such as a Web site, used to send bulk SMS messages and/or to provide instructions for participants. In some cases, additional infrastructure is needed to support a scenario where SMS is used in conjunction with another mobile technology (e.g., WAP), or a scenario where SMS learning is integrated with another e-learning approach to become part of a blended learning model.

Design and Pedagogy

Design and pedagogy describes the context for which the scenario was designed and developed, including its activities and expected learning outcomes. The framework in Figure 2 presents the general design contexts and the pedagogical aspects of m-learning.

Participants

The *participants* might be learners (university students, adult learners, or the general public) and teachers (faculty,

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. SMS learning scenarios: Design contexts and pedagogical aspects

SMS Scenarios: Design Contexts (Roibas, 2002; Bollen, Eimler, & Hoppe, 2004; Kadirire, 2005; Pincas, 2004; Colley & Stead, 2003; McMillan & Keough, 2005)	**SMS scenarios: Pedagogical Aspects (Pincas, 2004; Singh, 2003)**
• Supporting both independent & collaborative learning • Supporting "just-in time learning" • Supporting content delivery in a condensed format • Supporting multiple learners' learning styles	• The urgency of user needs • The ownership of initiative • The mobility of setting • The interactivity of process • The situated-ness of needs • The integration of content

administrators). Participants' background, perceptions, attitudes and priorities play a critical role in the successful adoption of a scenario where they have a stakeholder role (Barker, Krull, & Mallinson, 2005).

TYPES OF SMS-BASED SCENARIOS FOR MOBILE LEARNING

Scenario descriptions were extracted from the literature on mobile learning (2002-2005). These include cases from Europe (UK, Ireland, Greece, Germany, Finland), Asia (Japan, Thailand, Malaysia), Africa (South Africa), Australia and New Zealand. The scenarios were classified based on their context, content, and orientation (Figure 3).

Two major categories were identified, each comprising five scenario types: *learning* (delivery of new content, test and quizzes, learning for revision, simulation-based learning, collaborative learning), and *learning support* (student support, communication, teacher support, blogs, Q&A sessions). Examples that illustrate each type are presented as follows, providing details about the participants, the educational setting, and the required additional technology.

Figure 3. Types of SMS learning scenarios

Learning	Learning support
Delivery of new content	Student support
Tests and quizzes	Communication
Learning for revision	Teacher support
Simulation based learning	Blogs
Collaborative learning	Q & A sessions

Delivery of New Content

All examples in this category refer to learning a foreign language. The participants are learners—there is no interaction with teachers. The SMS server needs to be able to handle registrations and store content.

The InLET project (Pincas, 2004) provided language support for international tourists attending the Olympic games in Athens. Tourists who subscribed to the service received SMS messages with short phrases in Greek. They could also request and receive a translation from or into Greek of a commonly used phrase.

Another example is the structured short course in English delivered to working learners in Hong Kong (Song & Fox, 2005). New words and expressions were sent to learners on a regular basis, following a predefined sequence of learning tasks. The course material was also available on the Web. Similar scenarios were implemented in Japan (Thorton & Houser, 2005) and in Australia (Levy & Kennedy, 2005).

Commercial mobile services for language support are also available (Chinnery, 2006). Munro (2005) describes a commercial application, which delivers a pair of learning objects to a paid customer: a text-based object using SMS, and a sound object via podcast to an MP3 player or a smartphone.

Tests and Quizzes

Tests and quizzes are used for formal learning and for self-assessment. Tests are typically conducted in controlled conditions. Quizzes are used in class or as a supplementary homework activity. Receiving feedback ("the score") is an important feature of a test or a quiz.

Tretiakov and Kinshuk (2005) describe a scenario where students in class were given a quiz and then had to submit their answer via SMS. The quiz question could be "multiple choice," "fill in the blank" or "matching lists." Students could work individually or share a mobile phone to send the answer as a group. Individual feedback was accessible on a Web site. Very similar is the setting described in Iliescu and Hines (2005), which also expanded the range of question types to include free text questions.

The example found in Capuano et al. (2004) offers functionality similar to the scenarios above. However there is no need for an additional platform (the Web), as all communication is based on SMS: the question text is sent to the student who responds with the answer and receives feedback. In the example the SMS service is part of a larger, multi-channel, integrated platform.

Finally, a formal testing experiment is reported in Whattananarong (2004). Students sent test answers via SMS after being shown or read the questions in the classroom. There is immediate feedback.

Learning for Revision

Mellow (2005) describes the study platform StudyTXT, which was implemented to facilitate a "flash card" scenario at a New Zealand university: students studying sports medicine could download content relevant to a topic of their choice. They could see the list of available topics on a Web site. A similar use of the same platform is proposed in Petrova (2007).

A different approach is adopted in the scenario described in Riordan and Traxler (2003): rather than being given the option to request a revision snippet, students identified as being at risk of failing a test were sent revision tips and directions for further study by their lecturers.

Simulation-Based Learning

Cheung (2004) describes a series of classroom experiments involving simulation gaming in a postgraduate microeconomics class. Students were given the game plan and had to submit responses related to the game via SMS. All responses were transmitted from the mobile network over the Internet to the teacher at their workstation, who then used specialised software to generate individualised return messages. These were broadcast back to students, simulating direct interaction among students.

Another participatory simulation game is described in Lonsdale, Baber, and Sharples (2004). Participants were involved in a role play (as "water droplets") and were sent "entry" and "exit" messages (commands) directing them to perform certain actions in the physical classroom environment. As the simulation game was run under Java, the text messages were sent from a mobile phone connected to the PC running the game.

Collaborative Learning

Bollen, Eimler, and Hoppe (2004) describe a scenario for a constructive discussion implementing a decision support system ("Cool Modes") in a literature class. Students were assigned specific roles and sent text messages related to the role from a PDA (mimicking SMS). The messages were delivered via the internal wireless LAN to a MySQL database on the teacher's PC. The teacher could query the database and display the discussion threads to the class on an electronic whiteboard.

A similar scenario is presented in Kadirire (2005): while listening to a presenter, students used mobile phones to text their comments which were stored in a database and then subsequently organised and formatted to be displayed to the class. The presenter could give feedback based on the comments.

Student Support

In most instances of this scenario (Mohhamad & Norhayati, 2003; Stone, 2004a; 2004b; Nonyongo, Mabusela, & Monene, 2005), assistance is offered to a class or a cohort of students informing them about changes in schedules, or reminding them about deadlines and other events. A Web accessible database might be used to register users for the service.

Communication

Seppala and Alamaki (2002) describe an integrated scenario where voice communication, SMS messaging and a WAP gateway were used to create a media-rich environment for communications between instructors and trainee teachers. Text messages were used to transmit general information among all participants.

Teacher Support

Silander and Rytkonen (2005) provide an example for meeting teachers' pedagogical needs. SMS was used to assist lecturers as part of an intelligent tutoring system (Alykko). Text messages were stored in a Web accessible format and were used to record learning logs (similar to blogs), to send instructions to students and to respond to questions (similar to Q&A sessions).

Blogs

The blog scenario presented in Divitini, Haugalokken, and Morken (2005) included SMS as part of a learning management system (LMS). It enabled students doing their teaching practice to upload entries to a shared community blog and to individual student blogs. The blogs were accessible for viewing on a Web server.

Questions and Answers (Q&A)

This scenario involves a general tool available to all students on and off-campus, which collects questions and displays them on a Web site (anonimized). Answers from staff are also displayed (Ng'ambi, 2005).

SMS SCENARIOS: SUMMARY

Applying the design contexts framework (Figure 2, left-hand column) it can be concluded that:

- SMS allows for the development of both independent and collaborative learning models. In a significant number of examples a mixed model is implemented—where all participants can benefit from individual experiences
- Multiple learning styles might be supported in cases where SMS is integrated into a larger LMS or VLE (virtual learning environment).
- As a technology, SMS is especially well suited for "just-in-time learning;" however, if integrated with other technologies, technology access issues might interfere with the process (e.g., learners on the move might have access to a mobile phone but not to the Web).
- SMS content always needs to be condensed due to the limitations of the technology (160 characters in one text message).

Analysing the pedagogical aspects of SMS learning (Figure 2, right-hand column) it can be seen that:

- Learner support scenarios are especially well tailored to satisfy urgent needs as most are based on a pull approach (the learner owns the initiative). Some of the learning scenarios are also pull-based (e.g., the flash card scenario) and rely on the learner to realise their need and engage in an m-learning activity.
- While the initiative in some cases belongs entirely to the teacher (e.g., sending tips to "weak" students), a group of scenarios supports both push and pull approaches (e.g., scenarios which augment classroom activities.)
- All scenarios except one use mobile and wireless networks.
- Most scenarios are highly interactive, often using auxiliary infrastructure (e.g., for displaying results, or for providing detailed instructions).
- The enhancement of SMS through adding other technologies to the interactions interferes with the mobility of the learner (e.g., a mobile learner might not have mobile access to all technologies such as the Web, involved in an integrated scenario). The cost of integration needs to be balanced with the benefits derived from the use of other communication channels.
- In most scenarios the level of integration between content and situation is very high (e.g., discussions, Q&A sessions, simulations).

FUTURE TRENDS

SMS has already evolved into enhanced services, which are capable of transferring images and animation (EMS, MMS). Multimedia messaging offers an integrated platform and thus eliminates the need of embedding SMS into environments based on other technologies. MMS and EMS allow the learner to be truly mobile, as users with smartphones can have access to the mobile Internet and to the Web. Some of the SMS scenarios may migrate to MMS (one example is the study of a foreign language). New types of personal learning are expected to evolve around these technologies, and pedagogical models will need to be developed and evaluated alongside with the development of specially organised and formatted content. However, SMS scenarios will continue to be attractive in environments where cost of using more sophisticated or advanced technologies might be prohibitive.

CONCLUSION

This article extensively reviews the literature on mobile learning models using SMS. Based on the results of the review, it classifies SMS-learning scenarios into two major categories and provides examples to illustrate the types identified under each category. A framework of design contexts and pedagogical aspects is used to summarise the findings. Future trends including multimedia and MMS scenarios are also discussed.

REFRENCES

Attewell, J. (2005, October). From research and development to mobile learning: Tools for education and training providers and their learners. In *Proceedings of the 4th World Conference on Mobile Learning (MLEARN05).* Retrieved March 23, 2006, from www.mlearn.org.za/CD/papers/Attewell.pdf

Barker, A., Krull, G., & Mallinson, B. (2005). A proposed theoretical model for m-learning adoption in developing countries. In *Proceedings of the 4th World Conference on mLearning* (Paper 14).

Berth, M. (2005). Adaptive ethnography: Methodologies for the study of mobile learning in youth culture. Paper presented at *Seeing, Learning, Understanding the Mobile Age*, Budapest, Hungary. Retrieved January 20, 2006, from http://www.fil.hu/mobil/2005/Berth.pdf

Bollen, L., Eimler, S., & Hoppe, U. (2004). SMS-based discussions- technology enhanced collaboration for a literature course. In J. Roschelle, T.-W. Chan, Kinshuk, & S. J. H. Yang (Eds.), *Proceedings of the Second International Workshop on Wireless and Mobile Technologies in Education* (pp. 209-210).

Brown, T. (2005). Towards a model for m-learning in Africa. *International Journal on E-Learning*, *4*(3), 299-315.

Capuano, N., Gaetta, M., Miranda, S., & Pappacena , L. (2004). A system for adaptive platform independent mobile learning. In J. Attewell & C. Smith (Eds.), *Mobile Learning Anytime Everywhere: A Book of Papers from MLEARN 2004* (pp. 53-56).

Cheung, S. (2004). Fun and games with mobile phones: SMS messaging in microeconomic experiments. In R. Atkinson, C. McBeath, D. Jonas-Dwyer, & R. Phillips (Eds.), *Beyond the Comfort Zone: Proceedings of the 21st Australasian Society for Computers in Learning in Tertiary Education* (pp. 180-183).

Chinnery, G. (2006). Going to the MALL: Mobile assisted language learning. *Learning Language & Technology, 10*(1), 9-16

Colley, J., & Stead, G. (2003). Take a bite: Producing learning materials for mobile devices. In J. Attewell & C. Smith (Eds.), *Learning with Mobile Devices Research and Development: A Book of Papers from the 2nd World Conference on Mobile Learning (MLEARN 2003)* (pp. 43-46).

Divitini, M., Haugalokken, O., & Morken, E.M. (2005). Blog to support learning in the field: Lessons learned from a fiasco. In *Proceedings of the Fifth International Conference on Advanced Learning Technologies* (pp. 219-221).

Evans, D., & Taylor, J. (2004). The role of user scenarios as the central piece of the development jigsaw puzzle. In J. Attewell & C. Smith (Eds.), *Mobile Learning Anytime Everywhere: A Book of Papers from the 3rd World Conference on Mobile Learning (MLEARN 2004)* (pp. 63-66).

Finn, M. (2004). The handheld classroom: Educational implications of mobile computing. *Australian Journal of Emerging Technologies and Society*, *2*(10), 21-35.

Garner, I., Francis, J., & Wales, K. (2002). An evaluation of the implementation of a short messaging system (SMS) to support undergraduate students. In *Proceedings of the European Workshop on Mobile and Contextual Learning* (pp. 15-18). Birmingham, UK.

Grinter, R. E., & Eldridge, M. (2003). Wan2tlk?: Everyday text messaging. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 441-448).

Iliescu, D., & Hines, E. (2005). WES: The SMS based student feedback, voting and notification system. *Interactions 59*(1), Article 2. Retrieved January 3, 2006, from http://www2.warwick.ac.uk/services/cap/resources/interactions/archive/issue25/iliescu/

Kadirire, J. (2005). The short message service (SMS) for schools/conferences. *Recent Research Developments in Learning Technologies, 2*, 856-859

Kukulska-Hulme, A., Evans, D., & Traxler, J. (2005). *Landscape study and mobile learning in the post-16 sector: Summary report*. Retrieved January 23, 2006, from http://www.jisc.ac.uk/uploaded_documents/SUMMARY%20FINAL%202005.doc

Leung, C.-H., & Chan, Y.-Y. (2003). Mobile learning: A new paradigm in electronic learning. In *Proceedings of the 3rd International Conference on Advanced Learning Technologies* (pp. 76-80).

Levy, M., & Kennedy, C. (2005). Learning Italian via mobile SMS. In A. Kukulska -Hulme & J. Traxler (Eds.), *Mobile Learning: A Handbook for Educators and Trainers*. London: Taylor & Francis.

Lonsdale, P., Baber, C., & Sharples, M. (2004). Engaging learners with everyday technology: A participatory simulation using mobile phones. In S. Brewster & M. D. Dunlop (Eds.), *Proceedings of the 6th International Mobile Human-Computer Interaction Symposium* (pp. 461-465).

McMillan, J., & Keough, M. (2005). Seven reasons why mLearning doesn't work. In *Proceedings of the 4th World Conference on mLearning* (Paper 44).

Mellow, P. (2005). The media generation: Maximise learning by getting mobile. In *Proceedings of the 2005 Conference of the Australasian Association for Computers in Learning in Tertiary Education* (pp. 469-476).

MMA. (2006). Portio research: Mobile messaging futures 2005-2010. Mobile Marketing Association. Retrieved January 23, 2006, from http://mmaglobal.com/modules/wfsection/article.php?articleid=71

Mohhamad, M. A., & Norhayati, A. (2003). A short message service for campus-wide information delivery. In *Proceedings* *of the Fourth National Conference on Telecommunication Technology* (pp. 216-221). Malaysia.

Munro, A. (2005). 5th digit language application. In *Proceedings of the 4th World Conference on mLearning* (Paper 49).

Ng'ambi, D. (2005). Mobile dynamic frequently asked questions (m-DFAQ) for student and learning support. In *Proceedings of the 4th World Conference on mLearning* (Paper 51).

Nonyongo, E., Mabusela, K., & Monene, V. (2005). Effectiveness of SMS communication between university and students. In *Proceedings of the 4th World Conference on mLearning* (Paper 53).

Petrova, K. (2007). Mobile learning as a mobile business application. *International Journal of Innovation and Learning, 4*(1), 1-13.

Petrova, K., & Sutedjo, Y. (2004). Just-in-time learning: Ready for SMS? In Kinshuk, D. Samson, & P. Isaias (Eds.), *Proceedings of the International Conference on Cognition and Exploratory Learning in the Digital Age* (pp. 495-498).

Pincas, A. (2004). Approaches to just-in-time learning with mobile phones: A case study of support for tourists' language needs. In J. Attewell & C. Smith (Eds.), *Mobile Learning Anytime Everywhere: A Book of Papers from the 3rd World Conference on Mobile Learning (MLEARN 2004)* (pp. 157-162).

Prensky, M. (2005). What can you learn from a cell phone? Anything! *Innovate, 1*(5). Retrieved January 4, 2006, from http://www.innovateonline.info/index.php?view=article&id=83

Quinn, C. (2000, Fall). mlearning: Mobile, wireless and in-your-pocket learning. *LineZine Magazine.* Retrieved January 23, 2006, from http://www.linezine.com/2.1/features/cqmmwiyp.htm

Riordan, B., & Traxler, J. (2003). *Supporting computing students at risk using blended technologies*. Paper presented at 4th Annual Conference of the Learning and Teaching Support Network for Information and Computer Sciences, Galway, Ireland.

Riordan, B., & Traxler, J. (2005). The use of targeted bulk SMS texting to enhance student support, inclusion and retention. In *Proceedings of the 2005 International Workshop on Wireless and Mobile Technologies in Education* (pp. 257-260).

Roibas, A. C. (2002). *Designing scenarios of m-learning*. Paper presented at the Knowledge Management Workshop 2002, Multimedia University, Malaysia.

Seng, J.-L., & Lin, S. (2004). A mobility and knowledge-centric e-learning application design method. *International Journal of Innovation and Learning*, *1*(3), 293-311.

Seppala, P., & Alamaki, H. (2002). Mobile learning and mobility in teacher training. In M. Milrad, U. Hoppe, & Kinshuk (Eds.), *Proceedings of the 2002 International Workshop on Mobile technologies in Education* (pp. 130-135).

Silander, P., & Rytkonen, A. (2005). An intelligent mobile tutoring tool enabling individualization of students' learning processes. In *Proceedings of the 4th World Conference on mLearning* (Paper 59).

Singh, H. (2003). Leveraging mobile and wireless Internet. Retrieved January 3, 2006, from http://www.learningcircuits.org/2003/sep2003/singh.htm

Song, Y., & Fox, R. (2005). Integrating m-technology into Web-based ESL vocabulary learning for working adult learners. In *Proceedings of the 2005 International Workshop on Wireless and Mobile Technologies in Education* (pp. 154-163).

Stone, A. (2004a). Mobile scaffolding: An experiment in using SMS text messaging to support first year university students. In K. Kinshuk et al. (Eds.), *Proceedings of the Fourth International Conference on Advanced Learning Technologies* (pp. 405-409).

Stone, A. (2004b). Blended learning, mobility and retention: Supporting first-year university students with appropriate technology. In J. Attewell & C. Smith (Eds.), *Mobile Learning Anytime Everywhere: A Book of Papers from the 3rd World Conference on Mobile Learning (MLEARN 2004)* (pp. 183-185).

Stone, A., & Briggs, J. (2002). ITZ GD 2 TXT: How to use SMS effectively in m-learning. *Proceedings of the European Workshop on Mobile and Contextual Learning* (pp. 11-14).

Stone, A., Briggs, J., & Smith, C. (2002). SMS and interactivity: Some results from the field and its implication on effective use of mobile telephony for education. In M. Milrad, U. Hoppe, & Kinshuk (Eds.), *Proceedings of the 2002 International Workshop on Mobile Technologies in Education* (pp. 147-151).

Thorton, P., & Houser, C. (2005). Using mobile phones in English education in Japan. *Journal of Computer Assisted Learning*, *21*(3), 217-228.

Tretiakov, A., & Kinshuk, K. (2005). Creating a pervasive testing environment by using SMS messaging. In *Proceedings of the 2005 International Workshop on Wireless and Mobile technologies* (pp. 62-66).

Trifonova, A. (2003). Mobile learning: Review of the literature. University of Trento. Retrieved December 15, 2005, from http://eprints.biblio.unitn.it/archive/00000359/01/009.pdf

Vavoula, N., & Sharples, M. (2002). KLeOS: A personal, mobile, knowledge and learning organisation system. In M. Milrad, U. Hoppe, & Kinshuk (Eds.), *Proceedings of the 2002 International Workshop on Mobile Technologies in Education* (pp. 152-156).

Wagner, E. D. (2005). Enabling mobile learning. *Educause Review*, *40*(3), 40-53.

Whattananarong, K. (2004, September). *An experiment in the use of mobile phones for testing*. Paper presented at the International Conference on Making Educational Reform happen: Learning from the Asian Experience and Comparative Perspectives, Bangkok, Thailand.

KEY TERMS

Blended Learning: A learning paradigm where multiple pedagogical approaches are used.

Blogging: Periodic publishing on the Web in a specially designed space for collaborative writing (Web log).

E-Learning: Learning facilitated by information and communication technologies ("electronic learning").

EMS: Enhanced message service.

Flash Card Learning: A method of learning by memorising, often used in studying disciplines such as medicine or law. A flash card has two parts—a question part and an answer part.

Just-In-Time Learning: A learning model where learners acquire knowledge as the need arises

LMS: Learning management system

MMS: Multimedia message service

Smartphone: A mobile phone, which can connect to the Web via a browser and support email.

SMS: Short message service

WAP: Wireless application protocol

VLE: Virtual learning environment

Snapshot Assessment of Asia Pacific BWA Business Scenario

Chin Chin Wong
British Telecommunications (Asian Research Center), Malaysia

Chor Min Tan
British Telecommunications (Asian Research Centre), Malaysia

Pang Leang Hiew
British Telecommunications (Asian Research Center), Malaysia

INTRODUCTION

The world is moving forward at a soaring rate. Within this change wireless technology is rapidly evolving and is playing an increasing role in the lives of people throughout the world. The people of today demand hassle-free and compact products, which can be used at anytime and anywhere with "always-best-connected" network solutions (Wong, Tan, & Hiew, 2005b). Wireless technology is a possible solution to meeting the immediate needs of society in the case of high-speed data delivery.

This article is devoted to assessing the deployment of wireless networks in the Asia-Pacific region, with special focus on the existing Wi-Fi and the emerging WiMAX solutions. Further, the overviews of wireless technologies as used in different business environments are given. The article is categorized into two sections: logistics, and retail and distribution. Each section discusses an example of a wireless solution adopted in the Asia-Pacific region.

This article offers a compilation of wireless solutions in the Asia-Pacific in order to map possible future scenarios on the use of wireless technologies in this region. The article serves as a foundation for further studies concerning the use of wireless technologies to improve quality of life.

BACKGROUND

The wireless local area network (WLAN) based on the IEEE 802.11 family of standards has demonstrated great efficiency and received positive responses for delivering broadband services. On the other hand, IEEE 802.16 (WiMAX), is a new wireless standard for broadband wireless access (BWA). WiMAX,[1] an acronym that stands for Worldwide Interoperability for Microwave Access, is a certification mark for products that pass conformity and interoperability tests for the IEEE 802.16 standards (Marks, 2006). It further extends the performance of IEEE 802.11 (Wi-Fi[2]) in terms of capacity, coverage range, quality of service (QoS), and mobility (with 802.11e-2005).

Technological revolution in communications is taking place in logistics. The industry is advancing significantly, adopting wireless technologies to secure its assets and improve services. The benefits accrued using wireless solutions in logistics include lowered insurance premiums, instant notification of security breach, flexible and secure handling of high-security cargo by authenticated personnel, data access through a wide range of mobile devices, and so forth.

In addition, wireless solutions have the fastest returns on investment in the back office and supply chain functions of retail environments. Wireless applications in retail and distribution make workers more productive, streamline operations, help goods flow faster, and provide access to real-time data and inventory. As a result, productivity increases with the reduction in errors, which ultimately improves the customer's experience. Businesses in other sectors have embraced the information revolution to reduce costs and improve productivity (Frist & Clinton, 2004). They use information technologies not as an end, but as a means to innovate and improve.

Problem Description: Logistics Industry

Today transportation companies are experiencing unprecedented upheaval (Baracoda, 2005). Amid growing customer demands and soaring costs, the logistics industry struggles to develop successful business models that can drive profitable results and achieve customer loyalty. Managing logistics business in the Asia-Pacific is very challenging and highly complex due to multiple countries, currencies, languages, and customs; varying technologies and logistics infrastructure; and multi-modal transportation (The Logistics Institute-Asia Pacific, 2002). Other problems faced by the logistics industry include security breaches, theft, high insurance premiums, and inability to track goods delivery in real time.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Retail and Distribution

The common problems faced in retail and distribution include the time taken in filling out and sending order forms, as well as printing of customers' orders and receipts, failure to provide one-to-one effective marketing due to the inability to access customer and product data at all times, the hassle to retrieve an up-to-date product catalog, and so forth. The increasing general enthusiasm on mobile technologies such as Bluetooth and radio frequency identification (RFID) has a positive effect on the acceptance of new mobile applications and services in retail and distribution (Ondrus & Pigneur, 2004). This would explain the reasons why wireless point of sale (POS) solutions are adopted by retailers. Wireless payments and ticketing are becoming a new trend for quick-service-oriented industries such as toll booths (e.g., Smart Tag in Malaysia).

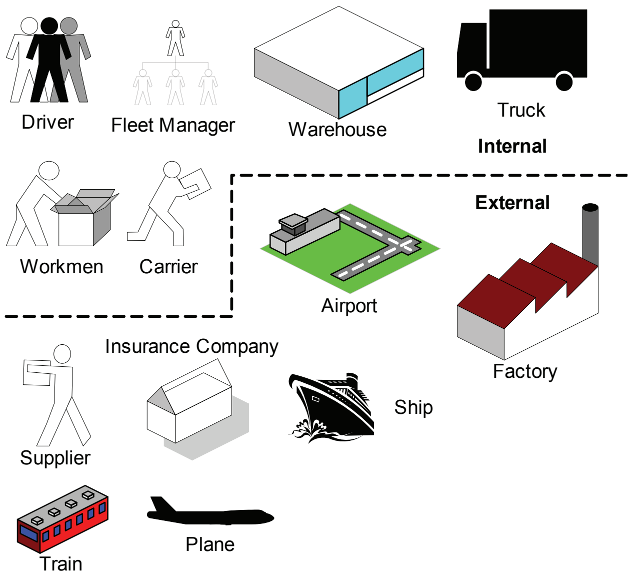
VIEWS OF ENVIRONMENT AND PROCESSES: LOGISTICS INDUSTRY

The cargo transportation industry is advancing significantly adopting wireless technologies to secure its assets and improve services (Nithyasree, 2005). Figure 1 shows the business environment of wireless applications used in the industry.

Uses of Wireless Solutions in Logistics Industry

- **Asset/ Cargo Tracking System:** A satellite-based vehicle tracking system using global positioning system (GPS) with satellite communications, geofencing[3], and cellular communication technologies allows fleet managers to remotely monitor, track, and communicate with their drivers in real time (Nithyasree, 2005).

Figure 1. Business environment for the logistics industry



- **Electronic Seals and RFID:** RFID technology is effectively utilized in the shipping and railroad industries alike. Electronic tracking tags and seals attached to a rail or ship create a WLAN that automatically informs the driver or a central control station of a broken seal (Nithyasree, 2005). These tags can also send vital information about the shipments such as the current status, whether tampered prior to destination, and so forth.

Example of Cargo Tracking System

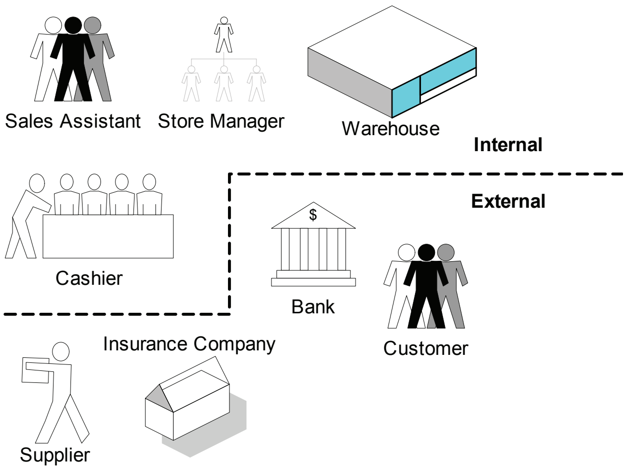
When the workmen pack goods to be delivered onto a truck, the fleet manager checks his personal digital assistant (PDA) for a list of guards on duty. He can see on his PDA the whereabouts of the security guards, and he makes sure that there is no sign of intrusion. Elsewhere, at a seaport, another fleet manager checks his PDA for information on each container, including its physical location based on GPS, parameters such as temperature and humidity, and whether there is any sign of intrusion. The information gathered can be connected to centralized databases. A service-oriented infrastructure allows the staffs to instantly share information. At the same time, a customer checks the location of his goods using his laptop at a hotspot (Wi-Fi). He is pleased that the goods will arrive on time. Once the goods are safely delivered to the customer, the driver enters details into his PDA to notify the fleet managers instantly. An example of a cargo tracking system deployed in the Asia-Pacific region is Kwikfleet (http://www.kwikfleet.com/).

Kwikfleet is a Malaysian company offering products and services either fully or jointly developed in Malaysia. With mobile data terminals (MDT), ruggedized portable computers, and wireless modems in their vehicles, fleet managers and their drivers in the field can take advantage of two-way computer-aided dispatching to stay connected while maintaining optimized scheduling and lowest time to destinations through advanced matching and dispatch algorithms. On-the-fly route planning technology will allow dynamic route planning algorithms to be run remotely on the MDT or locally on the intelligent vehicle location system server to serve portable data terminals (Kwikfleet, 2005). By using a geographic information system (GIS) and GPS, fleet managers are able to track a vehicle's location, speed, route traveled, as well as fuel level and so forth.

Business Processes

The logistics industry-related business processes involved in the cargo tracking system example are:

Figure 2. Business environment for retail and distribution



- security monitoring,
- asset/goods management,
- logistics personnel communications, and
- route selection.

Retail and Distribution

Retailers are using employee-activated wireless handheld devices to update inventory processes and increase accuracy and efficiency in all areas of the supply chain. In the process, they are taking significant steps toward reducing human error, returning salespeople to the business of helping customers, and avoiding problems like out-of-stocks and overstocking in order to gain competitive advantage (Schwartz, 2002). Figure 2 shows the business environment of wireless applications used in retail and distribution today.

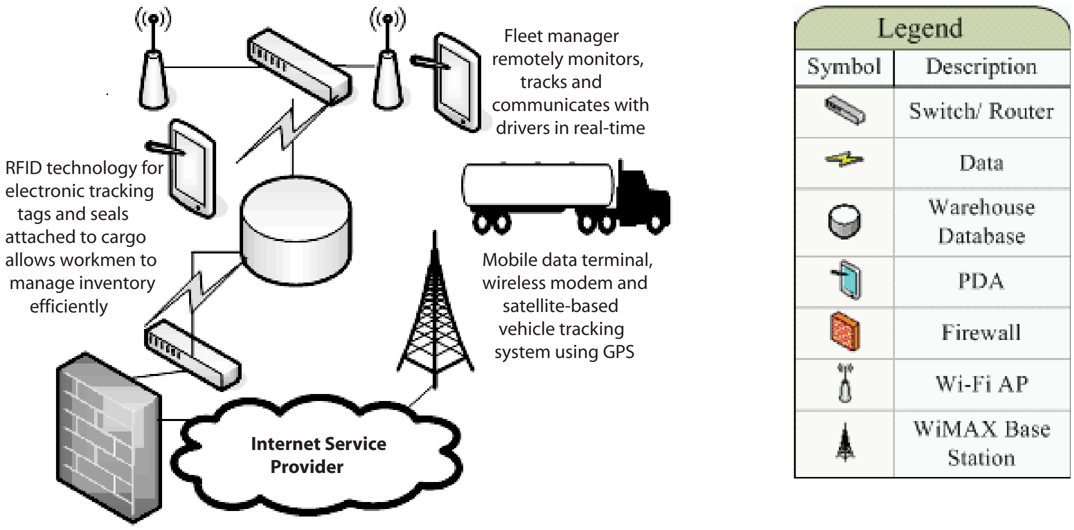
Uses of Wireless Solutions in Retail and Distribution Industry

- **Wireless POS:** Eliminating cables not only improves user convenience, productivity, and safety, but also cuts down on extra expenses over time. As cables age, they have to be replaced at a rate of two cables per terminal per year. Wireless POS provides better customer service by accommodating fluctuations in customer volumes and providing timely service. This can be achieved using Wi-Fi for WLAN connections, and subsequently using WiMAX or Mobile-Fi (based on IEEE 802.20) for connections between branches at different locations (Wu & Yallapragada, 2006).
- **Inventory Management and Replenishment:** A supplier typically would deliver a certain quantity of items, scratch out an invoice for the retail store manager to file away, and deliver a copy of that invoice back to the supplier's own accounts department for processing. Retailers frequently dispute bills submitted for payment because of pricing discrepancies, or charged back for unauthorized deliveries. Payments were slow and often incomplete. As a result of this tedious process, retailers suffered from inaccurate inventories. At the same time, suppliers were troubled by lengthy check-in times and high administrative costs, and struggled with remittance. By accessing information from inventory management by means of wireless access technologies, a store manager will be able to retrieve information about goods availability anywhere in real time and plan ahead to schedule purchases by notifying the suppliers. Suppliers will be able to respond almost instantly whether they are able to fulfill the order at a specific time and arrange for delivery.
- **Price Management:** If the retailer has a concern with pricing and price markdowns, a price management application with real-time access (mobile wireless) to the in-store computer has proven to improve item price accuracy (Pillar, 2003).

Example of Wireless Solutions for Retailers

At a corner of a supermarket, a customer scans a number of products at a kiosk while cruising through the store to check for prices and additional information. Elsewhere, a store manager accesses stock details from his portable device. The inventory management system prompts him to make a number of purchase orders to replenish a number of goods. He immediately sends purchase orders to suppliers. One of the suppliers responded that the goods could not be delivered on time. The store manager searches for other suppliers in order to stock up the goods. Customers are queuing up to make payment at the counters. One of the customers who is in a rush makes purchases online via his PDA and is happy that the goods will be delivered to his house within an hour. In the warehouse, a number of workmen place goods into a truck according to the list shown on their portable devices. An example of wireless solutions for retail and distribution industry deployed in the Asia-Pacific region is the SkyWire Wireless Automatic Data Capture Solution (http://www.skywire.com.au/). The solution aims at making stocktaking, price checking. and other retail applications more "hassle free." The improved functionality offered by real-time data flow between shop floor and back office systems helped a duty-free retail chain in Australia to deliver even better customer services and enjoy improved operational efficiencies (SkyWire, 2005).

Figure 3. Technical environment for wireless solutions in logistics industry


Fleet manager remotely monitors, tracks and communicates with drivers in real-time
RFID technology for electronic tracking tags and seals attached to cargo allows workmen to manage inventory efficiently
Mobile data terminal, wireless modem and satellite-based vehicle tracking system using GPS
Internet Service Provider
Legend
Symbol
Description
Switch/ Router
Data
Warehouse Database
PDA
Firewall
Wi-Fi AP
WiMAX Base Station

Business Processes

The retail and distribution business processes involved in the example of wireless solutions for retailers are:

- inventory management and replenishment,
- retailers and distributors negotiation, and
- price management.

FUTURE TRENDS: LOGISTICS INDUSTRY

Figure 3 proposes the devices required to deploy wireless solutions in the logistics industry. In this business scenario, PDAs can be used within the coverage of WLAN. Switches/routers are connected (wired) to a central database where fleet managers and workmen access timesheets, inventory, and so forth.

A firewall is required to protect the system from intrusion to ensure that confidential data is not tampered with. Access to the Internet is likely to be restricted by security policy. Drivers will be accessing the warehouse database using mobile WiMAX (802.16e-2005) or Mobile-Fi (802.20) which supports mobility.

RFID is an automatic identification method, relying on storing and remotely retrieving data using devices called RFID tags or transponders (Want, 2004). An RFID system may consist of several components: tags, tag readers, edge servers, middleware, and application software. The purpose of an RFID system is to enable data to be transmitted by a mobile device, called a tag, which is read by an RFID reader and processed according to the needs of a particular application. The data transmitted by the tag may provide identification or location information, or specifications about the product tagged, such as price, color, date of purchase, and so on.

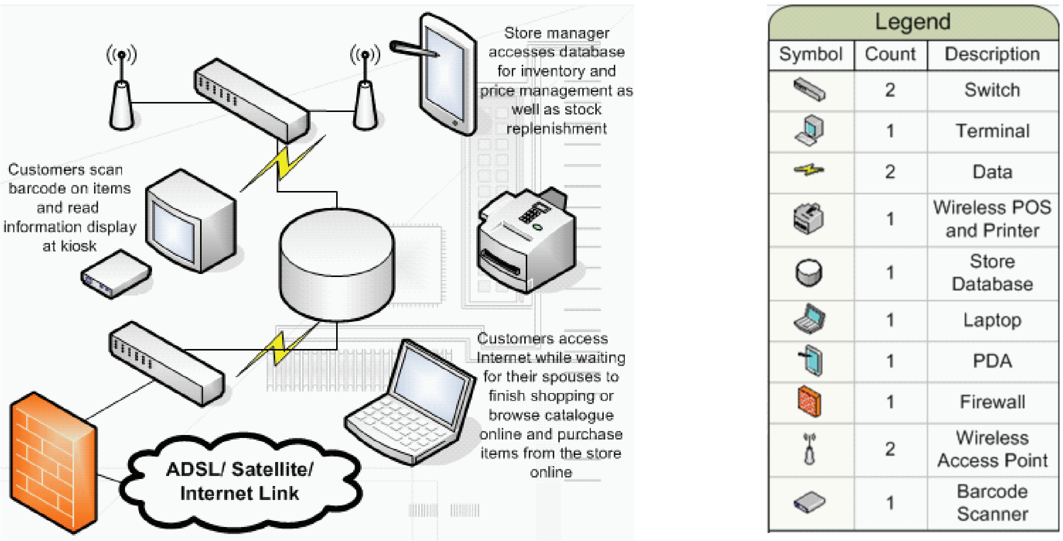
GPS is a satellite navigation system used for determining one's precise location and providing a highly accurate time reference almost anywhere on Earth or in Earth orbit (Beisner, Rudd, & Benner, 1996). The GPS system is divided into three segments: space, control, and user. The space segment comprises the GPS satellite constellation, whereas the control segment comprises ground stations around the world that are responsible for monitoring the flight paths of the GPS satellites, synchronizing the satellites' onboard atomic clocks, and uploading data for transmission by the satellites. The user segment consists of GPS receivers.

Retail and Distribution

Figure 4 depicts the devices required to deploy wireless solutions in retail and distribution. In this business scenario, PDAs and laptops can be used within the coverage area of WLAN. Switches/routers are connected (wired) to a central database where store managers and sales assistants access inventory. A firewall is required to protect the system from intrusion to ensure that confidential data is not tampered with. Access to the Internet is likely to be restricted by in-house security policy.

Customers may also scan barcodes found on items in the store to retrieve more information, for example, the number of items available, date of availability, item description, and so forth. Although wireless networking and computing is widely used in store operations, relatively few retailers have leveraged these investments with complementary wireless printing applications (Retail Biz, 2005).

Figure 4. Technical environment for wireless solutions in retail and distribution


Store manager accesses database for inventory and price management as well as stock replenishment
Customers scan barcode on items and read information display at kiosk
Customers access Internet while waiting for their spouses to finish shopping or browse catalogue online and purchase items from the store online
ADSL/ Satellite/ Internet Link
Legend

Symbol	Count	Description
	2	Switch
	1	Terminal
	2	Data
	1	Wireless POS and Printer
	1	Store Database
	1	Laptop
	1	PDA
	1	Firewall
	2	Wireless Access Point
	1	Barcode Scanner

Retailers can take advantage of wireless printing. Various wireless printing applications can lower total in-store printing expenses, provide total cost of ownership benefits compared with traditional printers, improve labor efficiency, reduce store operating expenses, improve safety measures, and increase customer satisfaction (Retail Biz, 2005).

Eliminating cables not only improves user convenience, productivity and safety but also cuts down on extra expenses significantly adding to the printer's cost of ownership over time (Retail Biz, 2005).

CONCLUSION

This article focuses on addressing one of the most widespread issues facing executives: aligning IT with business. The accomplishment of any major IT project is measured by the extent to which it is linked to business requirements, and demonstrably supports and enables the enterprise to reach its business goals (Wong, Tan, & Hiew, 2005a). This article has presented business scenarios in logistics as well as retail and distribution sectors. These analyses shape important techniques that may be exploited at various stages of defining enterprise architecture in order to derive characteristics of the architecture directly from high-level requirements of the business. In this study, this is achieved by examining business and technical environments, as well as related processes to enable successful deployment of wireless solutions in both of these industries. The technique has been used to help identify and understand business requirements, and hence to derive business requirements that the architecture development and ultimately the IT has to address. This helps to encourage the uptake of wireless technologies in the Asia-Pacific region.

REFERENCES

Baracoda. (2005). *When it comes to transportation Baracoda really delivers.* Retrieved September 26, 2005, from http://www.baracoda.com/baracoda/solutions/p_1.html

Beisner, H. M., Rudd, J. G., & Benner, R. H. (1996). Real-time APL prototype of a GPS system. *ACM SIGAPL APL Quote Quad, 26*(4), 31-39.

Frist, B., & Clinton, H. (2004). How to heal health care. *Washington Post,* (August 25), A17.

Kwikfleet. (2005). *About Kwikfleet.* Retrieved September 29, 2005, from http://www.kwikfleet.com/kwikfleet/index2.htm

Marks, R.B. (2006, January 15). *The IEEE 802.16 Working Group on Broadband Wireless Access Standards.* Retrieved January 18, 2006, from http://grouper.ieee.org/groups/802/16/

Nithyasree, M.G. (2005). *Wireless—Paramount in cargo security.* Retrieved September 27, 2005, from http://logistics.about.com/library/weekly/uc120602a.htm

Ondrus, J., & Pigneur, Y. (2004). Coupling mobile payments and CRM in the retail industry. *Proceedings of the IADIS International E-Commerce Conference,* Lisbon, Portugal.

Pillar, M. (2003). Where is wireless in retail? *Integrated Solutions Magazine*. Retrieved October 3, 2005, from http://www.ismretail.com/articles/

Retail Biz. (2005). High wired. *Retail Biz*. Retrieved Febuary 28, 2005, from http://www.ismretail.com/articles/

Schwartz, K. (2002). *Retail goes wireless.* Retrieved September 29, 2002, from http://www.kioskbusiness.com/Jan-Feb02/articles/article1.html

SkyWire. (2005). *Wireless mobile solution makes for "hassle-free" operations at downtown duty free stores throughout Australia.* Retrieved September 29, 2005, from http://www.skywire.com.au/show_this_item.php?pageId=123&secId=8&parentId=8&division=retail

The Logistics Institute-Asia Pacific. (2002). *The Logistics Institute-Asia Pacific launches centre of competence in optimization.* Retrieved September 26, 2002. from http://www.tliap.nus.edu.sg/tliap/Media_Events/E08Feb2002/E08Feb2002.aspx

Want, R. (2004). RFID: A key to automating everything. *Scientific American, 290*(1), 46-55.

Wong, C. C., Tan, C. M., & Hiew, P. L. (2005a, December). Business scenarios assessment in healthcare and education for 21st century networks in Asia Pacific. *Proceedings of the 8th International Conference on Enformatika, System Sciences and Engineering,* Krakow, Poland, 175-180.

Wong, C. C., Tan, C. M., & Hiew, P. L. (2005b, December). Early assessment of WLAN/ BWA exploitation opportunities in Asia Pacific. *Proceedings of the IADIS International Conference on E-Commerce,* Porto, Portugal, 434-438.

Wu, G., & Yallapragada, R. (2006). *IEEE 802.20 Mobile Broadband Wireless Access (MBWA).* Retrieved January 18, 2006, from http://grouper.ieee.org/groups/802/20/

KEY TERMS

Global Information System (GIS): Enables one to envision geographic aspects of a body of data. Basically, it allows query of a database and receives results in the form of map. A GIS can have many uses, for example, weather forecasting, sales analysis, population forecasting, land use planning, and so forth.

Global Positioning System (GPS): A "constellation" of 24 well-spaced satellites that orbit the Earth and make it possible for people with ground receivers to pinpoint their geographic location. The location accuracy is anywhere from 100 to 10 meters for most equipment. Accuracy can be pinpointed to within one meter with special military-approved equipment. GPS equipment is widely used in science and has now become sufficiently low cost so that almost anyone can own a GPS receiver.

Quality of Service (QoS): On the Internet and in other networks, QoS is the idea that transmission rates, error rates, and other characteristics can be measured, improved, and, to some extent, guaranteed in advance. QoS is of particular concern for the continuous transmission of high-bandwidth video and multimedia information. Transmitting this kind of content dependably is difficult in public networks using ordinary "best effort" protocols.

Radio Frequency Identification (RFID): A technology that incorporates the use of electromagnetic or electrostatic coupling in the radio frequency portion of the electromagnetic spectrum to uniquely identify an object, animal, or person. RFID is coming into increasing use in industry as an alternative to the bar code. The advantage of RFID is that it does not require direct contact or line-of-sight scanning.

Wireless Fidelity (Wi-Fi): A term for certain types of WLAN that use specifications in the 802.11 family. The term Wi-Fi was created by an organization called the Wi-Fi Alliance, which oversees tests that certify product interoperability. Wi-Fi has gained acceptance in many businesses, agencies, schools, and homes as an alternative to a wired LAN. Many airports, hotels, and fast-food facilities offer public access to Wi-Fi networks. These locations are known as hotspots.

Wireless Point of Sale (POS): A component of wireless telemetry, or machine-to-machine correspondence. Specifically, wireless POS is the ability to make a purchase using a credit or debit card in businesses that would use a wireless POS terminal, such as taxicabs, limos, home repair services (e.g., carpet cleaners or refrigerator repair), restaurants, or mobile merchants.

Worldwide Interoperability for Microwave Access (WiMAX): A wireless industry coalition whose members organized to advance IEEE 802.16 standards for BWA networks. WiMAX 802.16 technology is expected to enable multimedia applications with wireless connection and, with a range of up to 30 miles, enable networks to have a wireless last-mile solution.

ENDNOTES

[1] For further reading, access the WiMAX Forum at http://www.wimaxforum.org/

[2] For further reading, access the Wi-Fi Alliance at http://www.wi-fi.org/

[3] Restrict the movement of a vehicle or other object to within a specified area. The location of the vehicle is monitored by telemetry, and an alarm is raised if it goes outside that area.

Software Platforms for Mobile Programming

Khoo Wei Ju
Malaysia University of Science and Technology, Malaysia

K. Daniel Wong
Malaysia University of Science and Technology, Malaysia

INTRODUCTION

Java 2 Micro Edition (J2ME), .NET Compact Framework (.NET CF), and Active Server Pages .NET (ASP.NET) Mobile Controls are commonly used alternatives in mobile programming. They provide an environment for applications to run on mobile devices. However, they are different in many ways, such as supported mobile devices, architecture, and development. Hence, it is important for mobile application developers to understand the differences between them in order to choose the one that meets their requirement. Therefore, in this article we will discuss the general architecture of J2ME, .NET CF and ASP.NET Mobile Controls and compare the three alternatives.

BACKGROUND AND INTRODUCTION

Since the mid-1990s, the growth of wireless communications has led to the mushrooming of mobile devices in the market. Initially, the mobile devices were mainly cell phones with limited programmability. However, many analysts and company executives were worried that mobile phone sales would eventually slow down, prompting research and development into software suitable for cell phones (Grice & Charny, 2001). Hence, now, there is a rise of programmable mobile devices. Furthermore, programmable mobile devices these days include not just cell phones but smartphones, PDAs, and pocket PCs. There are three well-known alternatives in mobile programming for general-purpose applications: J2ME, .NET CF, and ASP.NET Mobile Controls.

J2ME is a version of Java that provides an application environment running on consumer devices and embedded devices. It targets machines with as little as 128KB of RAM (Tauber, 2001). J2ME consists of Java virtual machines (JVMs) and a set of standard Java application program interfaces (APIs) defined through the Java community process (JCP). J2ME can be used with different configurations and profiles, which provide specific information to a group of related devices. Configurations support the Java core APIs. Profiles are built on top of configurations to support device-specific features like networking and user interfaces. The J2ME is available in two main configurations: connected limited device configuration (CLDC) and connected device configuration (CDC). Figure 1 shows the hierarchical structure of J2ME.

.NET CF is a lightweight version of Microsoft's .NET framework. It provides an environment for executing client-side code and eXtensible Markup Language (XML) Web services to smart devices. It is compatible with C# and Visual Basic.NET (VB.NET), and it supports (.NET Compact Framework Team, 2005):

- Windows mobile (2000, 2002, 2003)-based pocket PC,
- Windows mobile-based smartphones, and
- embedded systems running Windows CE .NET 4.1 and later.

.NET CF consists of two main components: the development environment and the runtime environment. The development environment, known as smart device extensions (SDEs), is a Visual Studio .NET (VS.NET) 2003 project type that allows .NET CF applications to be developed rapidly by simply dragging appropriate controls into the application. The runtime environment is the common language runtime (CLR). The size of the CLR and relevant class libraries is smaller than 2MB, which is suitable for mobile devices. The architecture of .NET CF is shown in Figure 2.

Active server pages (ASPs) is Microsoft's server-side scripting technology. An active server page has an .asp extension, and it mixes HyperText Markup Language (HTML) and scripting code that can be written in VBScript or JavaScript. ASP is distributed with Microsoft's Internet information services (IIS) Web server, so most hosts using IIS will also offer ASP for dynamic Web programming. ASP.NET is the version of ASP that works with Microsoft's .NET Framework.

ASP.NET Mobile Controls was previously known as Microsoft mobile Internet toolkit (MMIT). It was renamed as ASP.NET Mobile Controls to reinforce the concept that it is a collection of ASP.NET controls designed for mobile applications. It extends the ASP.NET server-side technology to allow developers to develop applications for a variety of mobile devices. Executing on the IIS Web server, ASP.NET Mobile Controls allows Web applications to be accessed by

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Hierarchical structure of J2ME

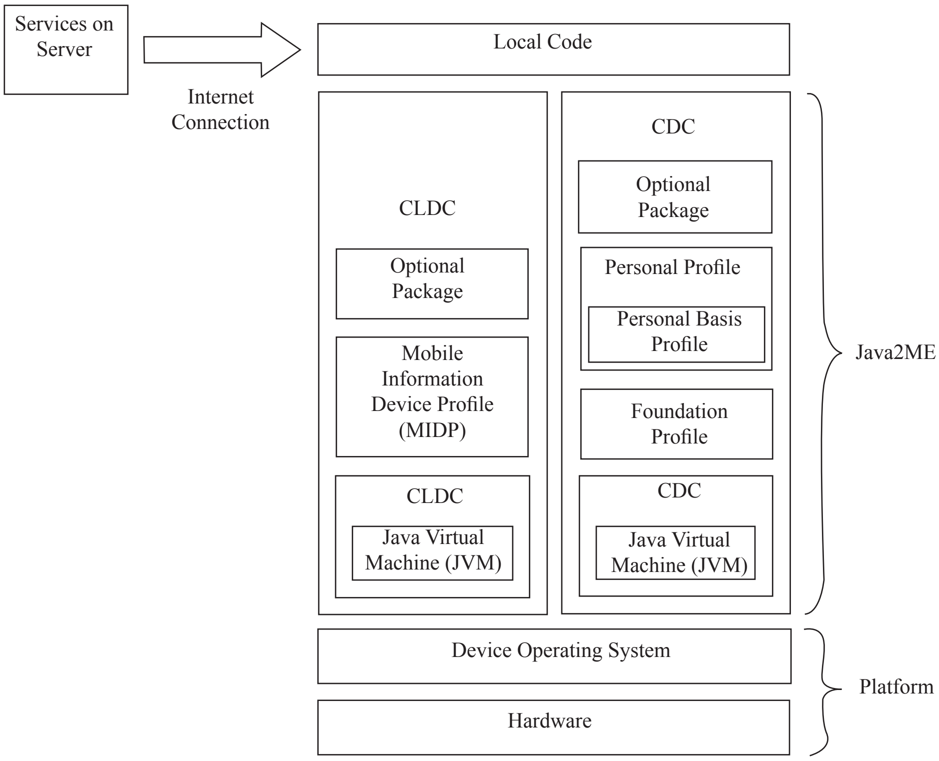
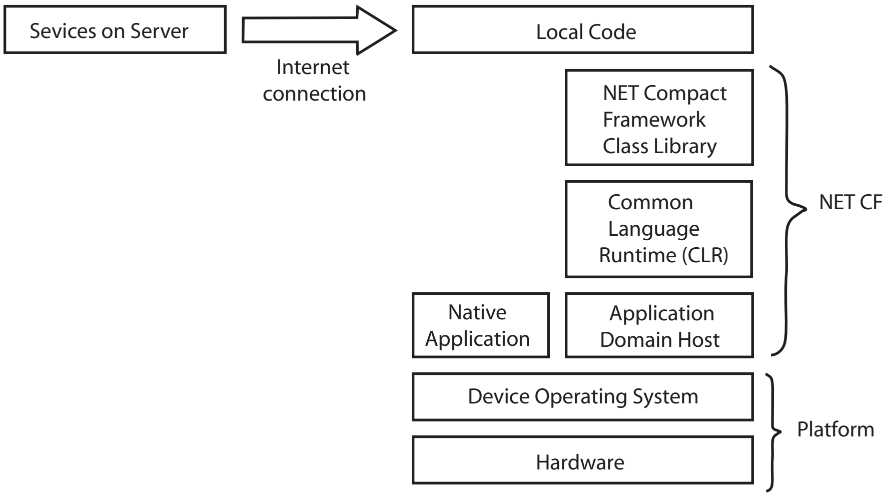
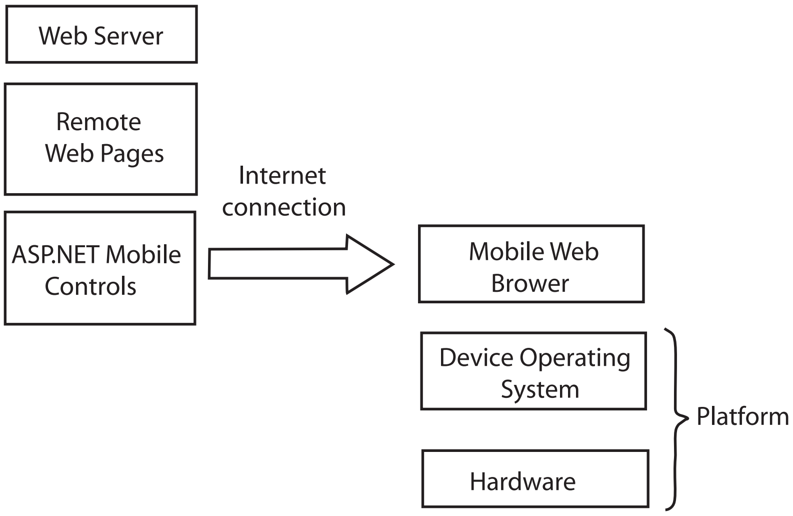

Services on Server
Internet Connection
Local Code
CLDC
Optional Package
Mobile Information Device Profile (MIDP)
CLDC
Java Virtual Machine (JVM)
CDC
Optional Package
Personal Profile
Personal Basis Profile
Foundation Profile
CDC
Java Virtual Machine (JVM)
Java2ME
Device Operating System
Hardware
Platform

Figure 2. .NET compact framework


Sevices on Server
Internet connection
Local Code
NET Compact Framework Class Library
Common Language Runtime (CLR)
Native Application
Application Domain Host
NET CF
Device Operating System
Hardware
Platform

almost any Internet-enabled mobile device. During runtime, it will automatically detect the device running the application. The application is then transformed into a form suitable for that device. This frees the developer to concentrate on the application logic and leaves the user interface rendering to the runtime (Lee, 2002a). Furthermore, it allows developers to visually drag and drop controls on forms aimed at mobile devices using VS.NET. The rest of the work, such as writing the proper markup language (e.g., Wireless Markup Language (WML), wireless application protocol (WAP)), is handled by the toolkit. The application development environment for ASP.NET Mobile Controls should be familiar to most ASP.NET programmers. Figure 3 shows the architecture of ASP.NET Mobile Controls.

Figure 3. ASP.NET mobile controls



FEATURE COMPARISON

J2ME, .NET CF, and ASP.NET Mobile Controls cannot be easily compared feature-to-feature because the analysis must include non-technology aspects such as market acceptance, development and testing tools, reach, standardization, and platform coherence. Besides, the final releases of J2ME's mobile information device profile (MIDP), personal basis profile, and personal profile are still in production (Sun, 2005). On the other hand, .NET CF is in the final stages of its beta tests. Nevertheless, a feature comparison, although limited, should still be useful.

Flexibility of Machine Control and Scope of Applications

Virtual Machines, Pointers, Native Features

In the .NET Compact Framework, the common language runtime (CLR) environment executes .NET's Microsoft Intermediate Language code. The CLR also offers support services, such as code verification, memory, and code security. The managed code is always translated into native machine code rather than interpreted. CLR supports interfaces and pointers. As for security policy, .NET CF grants full trust to all code (Microsoft, 2005b). The standard frameworks cover only a limited set of commonly used mobile device features. Other features are accessible via native methods. Besides, it is believed that .NET CF has better support for native methods than J2ME because Microsoft controls both .NET CF and the Windows operating system (Yuan, 2002).

With J2ME, Java source code is compiled into machine-independent byte code. The byte code is then interpreted by the Java virtual machine (JVM) during runtime. J2ME employs different versions of the JVM based on the needs of a particular situation. The configuration specifications define the characteristics of the J2ME virtual machines. In most cases, features of the JVM are removed to accommodate the needs of a configuration. The CDC runs on a C-virtual machine (CVM) that is fully compliant with the Java virtual machine specification. The CDC profile accommodates devices with as little as 512kB of memory, although it is really designed for platforms with about 2 MB of available memory (White & Hemphill, 2002). Sun provides a reference implementation of the CLDC specification that is based on the KVM, a small footprint of JVM that satisfies the CLDC requirements. However, products need not be based on KVM—any virtual machine that has the features required by the specification and can work within the resource restrictions of the CLDC environment can be used (Topley, 2002). Although JVM supports interfaces, it does not support pointers because it can result in unsafe code. The Java native interface (JNI), which allows access to native methods, can be used but only by CDC. For CLDC, the native features must be built into the runtime.

Consumer Applications, Multimedia, Gaming

.NET CF supports direct draw on canvas, double buffering, and device button remapping through its rich Windows Forms User Interface library. It also supports multimedia playback by using the native methods from Windows Media Player on Pocket PC (Yuan, 2002).

In J2ME, the mobile information device profile (MIDP) 2.0 for CLDC includes animation and game controls in the javax.microedition.lcdui.game package. Multimedia playback is supported via the Java media framework (JMF) on the CDC or the multimedia optional package for the CLDC. Many game developers prefer J2ME, because it is supported by a wider range of mobile platforms.

Development Support

Programming Languages

ASP.NET supports any language supported by the .NET Framework, including C, C++, C#, Visual Basic, and even Java. However, .NET CF currently supports only two major .NET languages: C# and VB.NET (Microsoft, 2005a). C# and VB.NET are standardized by EMCA and ISO/IEC. Hence, Microsoft has long been criticized for tightly controlling its technologies. However, the support of multiple standardized languages allow developers flexibility in programming in .NET CF.

J2ME only supports Java. Anyone can propose a Java specification request (JSR) to the Java community process (JCP) for a new platform extension. Unlike with the tightly controlled development of .NET compact framework and ASP.NET, it may appear that under a more free process like the JCP, developers have to spend much time understanding the features to make use of all extensions in the language. However, J2ME APIs undergo rigorous standardization processes to ensure wide industry support and minimum learning for developers.

Platforms

.NET CF supports high-end PDAs such as Windows pocket PCs, Windows smartphones, and embedded devices running on the Windows CE .NET platform (Microsoft, 2005a). Windows devices consist of only a small part of today's mobile device population.

With J2ME, most of the cell phone devices (Motorola iDEN, Nokia Symbian OS, and Qualcomm Brew platforms) and low-end PDAs (Palm OS and Real-Time OS platforms) have built-in Java support because Java allows developers to be productive across many mobile platforms.

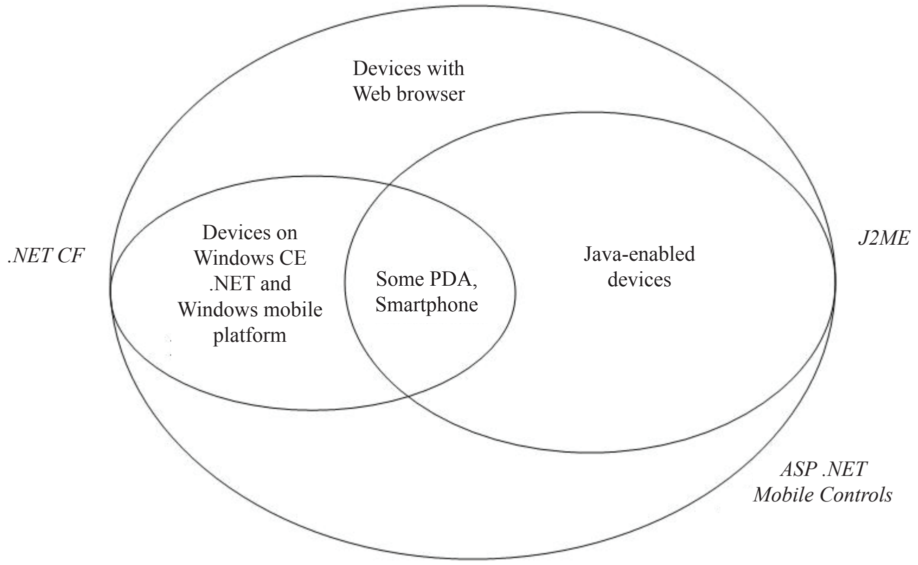
Because it is server-side based, placing minimal requirements on the client, ASP.NET is supported by the widest range of mobile devices, including all devices that support .NET CF, all devices that support J2ME, and more. However, each of these devices may present the output of the ASP. NET controls differently due to their different limitations and capabilities.

Development Tools

Regarding .NET CF and ASP.NET, Visual Studio .NET provides similar design interfaces for both mobile and non-mobile applications. It supports Web services integration and relational database access, and VS.NET is tightly integrated with Visio Enterprise Network Tools edition, which can generate C# or VB.NET code from UML (Unified Modeling Language) diagrams. Furthermore, VS.NET supports debugging on both emulators and real devices. However, VS.NET is not free.

Sun's J2ME Wireless ToolKit is a widely used MIDP development tool. Furthermore, command-line tools and vendor-specific toolkits are readily available. All major Java integrated development environments (IDEs) have J2ME modules or plug-ins. A big challenge for all J2ME IDEs is vendor software development kit (SDK) integration. Every

Figure 4. Devices supported by .NET CF, J2ME, and ASP.NET mobile controls



device vendor provides SDKs for their device emulators and proprietary J2ME extensions. The unified emulator interface (UEI) is designed to standardize the interfaces between IDEs and device SDKs. However, the UEI is available only through a Sun licensing program.

Miscellaneous

Specification Process

When a new technology emerges, Microsoft has the veto power to make decisions and make it available on .NET CF and/or ASP.NET. This saves time and effort. On the other hand, this also means that developers have no say on the specification process.

The Java community process (JCP) decides the new J2ME standard APIs, whereas Sun has veto power on only Java language specifications. The JCP develops all current J2ME configurations, profiles, and optional packages, so the specification process is arguably very lengthy and inefficient. However, some developers like this because more people are allowed to contribute and decide.

Gateways

There are technical difficulties in using .NET CF in mobile gateways because it was not designed to run lightweight application servers required in mobile gateways. Although Microsoft mobile information server (MIS) is a powerful gateway, messaging, and synchronization server, .NET CF lacks built-in APIs to interact with Microsoft MIS (Yuan, 2002).

For J2ME, the primary mobile service gateway product is from IBM. The Oracle9i wireless application server and Oracle J2ME SDKs provide gateway integration points for mobile devices to many other Oracle or third-party application servers (Yuan, 2002).

Additional Comparisons Between .NET CF and ASP.NET Mobile Controls

Server-Client Side

.NET CF uses client-side technology. Code is executed on the mobile device using just-in-time (JIT) compilation and native execution. ASP.NET uses server-side technology. Code is executed on the server, producing markup-language-based output such as HTML to be interpreted by a Web browser.

Web Server

For .NET CF, a Web server is not needed because code is executed directly on the device. For ASP.NET, a Web server (such as Microsoft IIS 5.0 or 6.0) that supports ASP.NET is required. The ASP.NET HTTP runtime is used to handle and process requests via a set of ASP.NET server controls.

Device Support

Only devices that have .NET CF runtime can execute programs written on .NET CF. On the other hand, since ASP.NET is server-side based, less processing is required on the client side compared to .NET CF. However, each of these devices may present the output of the ASP.NET controls differently due to their different limitations and capabilities.

Connectivity

For .NET CF, standalone applications can be created and installed on portable devices such as the pocket PC, pocket PC phone, and smartphone. By having the applications downloaded into such devices, the devices can either be connected or not connected. The XML or SQL Server 2000 Windows CE editions are used for local storage when

Table 1. Comparison summary between .NET CF, J2ME (CDC and CLDC) and ASP.NET Mobile Controls

	.NET Compact Framework	**J2ME Connected Device Configuration**	**J2ME Connected Limited Device Configuration**	**ASP.NET Mobile Controls**
Virtual Machine	Common Language Runtime (CLR)	Java Virtual Machine (JVM)		Common Language Runtime (CLR)
Portable Code	Intermediate Language (IL)	Byte code		Intermediate Language (IL)
Just-In-Time (JIT) Compiling	Yes	Yes		Yes
Garbage Collection	Yes	Yes		Yes
Portability	No	Yes		Yes

Table 1. continued

Cross-Language Integration	Yes	No		Yes
Standardized	EMCA, ISO/IEC	Yes		EMCA, ISO/IEC
Server-Client Side	Client side	Client side		Server side
Web Server	Not needed	Not needed		Needed
Device Support	Pocket PC, smartphone, Windows CE	General-purpose Java phone, smartphone, and PDA		Device independent
Connectivity	Standalone	Standalone		Connected
Market Focus	Enterprise	Enterprise	Consumer and enterprise	Consumer and enterprise
Language Support	VB.NET, C#	Java	Java	VB.NET, C#, C++, C, Java
Platforms	Pocket PC, Windows CE	Major mobile platforms except Palm OS	All mobile platforms	All mobile platforms
API Compatibility	Subset of .NET	Subset of J2SE plus standard optional packages	Partial compatibility with CDC with additional standard optional packages	Subset of .NET
Native APIs	Platform Invoke	JNI; device and OS specific	-	-
Coding and Development Tools	Smart Device Programming (SDP), Microsoft Visual Studio .NET	Command line, vendor SDKs, CodeWarrior, and WebSphere	Command line, vendor SDKs, all major Java IDEs	Microsoft Visual Studio .NET
Specification Process	Single company	Community	Community	Single company
Service Gateway	-	Run gateways as OSGi servlets; run gateway clients via vendor-specific SDKs	Run gateway clients via vendor-specific SDKs	-
Security Model	Simplified.NET model	Full Java security manager	Limited Java 2 model supplemented by OTA specification	Simplified.NET model
Client Installation	ActiveSync, Internet Explorer download	Sync, download	Formal OTA specification	
Lifecycle Management	-	OSGi for gateway apps, J2EE Client Provisioning Specification for generic clients	Included in OTA spec, works with J2EE Client Provisioning Specification	-
User Interface	Rich subset of Windows Forms	Rich subset of AWT (Abstract Windowing Toolkit), vendor-specific UI libraries	PDA Profile subset of AWT, vendor-specific UI libraries	Rich subset of Windows Forms
Mobile Database	SQL Server CE, Sybase iAnywhere Solutions(coming soon)	IBM DB2 Everyplace, iAnywhere Solutions, PointBase, Oracle9i Lite	Vendor-Specific relational implementation over RMS, Oracle SODA	SQL Server CE
Database Synchronization	Vendor specific	Vendor specific	Vendor specific	Vendor specific
XML API	Built into ADO.NET and other standard APIs	Third-party tools	Third-party tools	Built into ADO.NET and other standard APIs

Table 1. continued

E-Mail and PIM (Personal Information Manager)	Platform Invoke—Outlook APIs	PDA optional packages	PDA optional packages	-
Short Message Service (SMS)/Multimedia Messaging (MMS)	Platform Invoke—SMS/MMS	Wireless Messaging API (WMA)/WMA 2.0	Wireless Messaging API (WMA)/WMA 2.0	Simple Mail Transport Protocol (SMTP) and third party
Instant Messenger	Platform Invoke—Microsoft Network (MSN) and other IM client APIs	Third-party APIs for most IM clients including Jabber and Jxta	Third-party APIs for most IM clients including Jabber and Jxta	-
Enterprise Messaging	Platform Invoke—Microsoft Message Queuing (MSMQ)	Proprietary JMS (Java Message Service) APIs	JMS via third-party toolkits (e.g., WebSphere MQ Everyplace, iBus Mobile)	-
Cryptography	Third-party APIs	JCE (Java Cryptography Extension) and third-party libraries	Third-party libraries	Third-party APIs
Multimedia	Platform Invoke—Windows Media Player APIs	Subset of Java Media Framework (JMF)	Built into MIDP plus J2ME multimedia APIs	-
Game	Included Windows Forms UI	Direct draw on Canvas	GameCanvas support in MIDP	-
Location API	APIs provided by carriers	Location API	Location API	-

working off-line. With ASP.NET Mobile Controls, an HTTP connection is required to request an ASP.NET page that uses the Mobile Controls.

FUTURE TRENDS

In recent years, a strange trend can be seen in the design of mobile devices; they are getting bigger and bigger. They are gradually taking on more of the features of regular computers. We say this is a strange trend because mobile devices were originally meant to be stripped down, barebones devices with only the most useful features for mobile usage. Nevertheless, this trend is getting encouraging responses from users because the mobile devices are able to hold all the files they need to carry around. Due to the encouraging response from the users and the advances in technology, it is predicted that the trend will continue.

Besides, the sales of traditional PDAs have declined in the past few years (ETForecasts, 2003). This is because the PDA market is gradually being taken over by the smartphone, also known as the PDA phone. Users favor phones with computer features, such as storage capacity and clearer display (Kewney, 2005).

Currently, non-Microsoft operating systems like Symbian dominate the smartphone operating system market. As the sales and variety of smartphones are increasing worldwide, assuming Windows mobile platform keeps the same percentage of the market, the usage of Windows mobile platform will grow as well. Besides, given the past success of Microsoft to expand into new related markets, it is quite likely that they could increase their market share over the next few years, and there is much room for them to grow. Since Windows mobile platform will support only .NET CF (Java is not included), the growth of Windows mobile platform will directly lead to the increase in the use of .NET CF in mobile programming.

Furthermore, the current versions of .NET CF grant full trust to all code. However, the upcoming versions of .NET CF will offer a subset of the policy-driven, evidence-based code access security of the full .NET framework (Microsoft, 2005b). This is a good feature from the security point of view; however, it may be less convenient for developers compared to the current version.

Hopefully in the future, both Microsoft and Java applications can coexist in the same mobile device.

CONCLUSION

.NET CF and J2ME are both excellent platforms for developing smart clients for mobile commerce applications. J2ME has already gained a lot of industry support as the most favorable platform for developing mobile applications, and there are over 500,000 skilled Java developers around the world (Wishart, 2002). J2ME implements a modular design and is portable across a variety of devices. The platform provides balanced support for both enterprise and consumer applications. J2ME vendors offer excellent selections of mobile databases and gateway application server products.

On the other hand, .NET has the advantage over J2ME where it provides a single development platform and common coding practices based around Visual Studio. There are more than 1.5 million skilled developers worldwide (Wishart, 2002), and Microsoft has the largest tools and third-party developers program. The .NET CF platform focuses on enterprise applications with rich user interface, database, and XML Web services support. Hence, .NET CF is suitable for cash-rich customers with controlled mobile environments. However, .NET CF runs only on Windows-powered high-end PDAs. As a young platform, it currently lacks support for gateway servers and choices for mobile databases.

For the near future, the choice between .NET CF and J2ME is not so much a question of the desired platform features (both are excellent in this respect) as the targeted devices. In the short run, J2ME is supported by more devices than .NET CF. In the long run, most experts expect both platforms to coexist in all market sectors. Developers must choose the right tools and make them all work in heterogeneous environments. For example, J2ME clients would need to work with .NET backend servers and vice versa. So it would ultimately not come down to a choice between J2ME or .NET CF.

Both .NET CF and J2ME have advantages over ASP.NET Mobile Controls in supporting code that will execute on the device, and that can run in disconnected, connected, or occasionally connected modes. Therefore, for most enterprise mobile solutions, .NET CF and J2ME are more appropriate.

On the other hand, if browser-based applications are required, Microsoft ASP.NET Mobile Controls can be used to develop mobile Web applications that adapt their page rendering for a range of devices, such as micro-browsers on PDAs, smartphones, and WAP phones. ASP.NET Mobile Controls allows the developers to target the users they need to target, without worrying about the device they are using.

ACKNOWLEDGMENTS

The assistance of Lisa Tang in reviewing, and commenting on, a draft of this article is gratefully acknowledged.

REFERENCES

ETForecasts. (2003, June 16). *Smartphones have started to impact PDA sales.* Retrieved December 13, 2005, from http://www.etforecasts.com/pr/pr0603.htm

Faridi, M. (2003). *Beginning compact framework.* Retrieved from http://www.ilmservice.com/twincitiesnet/presentations/BeginningCF.NET.ppt

Grice, C., & Charny, B. (2001, February 2). *Wireless jungle still waiting for its king.* Retrieved from http://news.com.com/2100-1033-252009.html

Jagers, B. (2003). *Comparing file transfer and encryption performance of Java and .NET.* Retrieved from http://www.lore.ua.ac.be/Publications/pdf/Jagers2004.pdf

Kewney, G. (2005, February 8). *Landscape phones mark the resurgence of the PDA smartphone.* Retrieved December 13, 2005, from http://www.newswireless.net/index.cfm/article/1918

Lee, W. (2002a, December 2). *Developing mobile applications using the Microsoft Mobile Internet Toolkit.* Retrieved from http://www.devx.com/wireless/Article/10148

Lee, W. (2002b, November 18). *Announcing .NET Framework 1.1.* Retrieved from http://www.ondotnet.com/pub/a/dotnet/2002/11/18/everett.html

Leghari, N. (2003, December 17). *Tools and platforms: Choices for a mobile application developer.* Retrieved from http://weblogs.asp.net/nleghari/articles/mobiledeveloper.aspx

Microsoft. (2005a). *.NET Compact Framework.* Retrieved from http://msdn2.microsoft.com/en-us/library/f44bbwa1(en-us,vs.80).aspx

Microsoft. (2005b). *Security in the .NET Compact Framework.* Retrieved from http://msdn2.microsoft.com/en-us/library/13s3wxyw.aspx

Milroy, S. (2003, March 6). *.NET Compact Framework overview.* Retrieved from http://www.windowsitpro.com/Articles/Index.cfm?ArticleID=38314&DisplayTab=Article

NET Compact Framework Team. (2005, January 6). *.NET Compact Framework FAQ.* Retrieved December 6, 2005, from http://msdn.microsoft.com/smartclient/community/cffaq/default.aspx

Sun. (2005). *Java 2 Platform Micro Edition (J2ME).* Retrieved December 8, 2005, from http://java.sun.com/j2me/index.jsp

Tauber, D. A. (2001, August 3). *What's J2ME?* Retrieved from http://www.onjava.com/pub/a/onjava/2001/03/08/J2ME.html

Topley, K. (2002). *J2ME in a nutshell.* O'Reilly & Associates.

White, J.P., & Hemphill, D.A. (2002). *Java 2 Micro Edition.* Manning Publications.

Wishart, A. (2002, April 29). *Mobile development environments: .NET Contra J2ME.* Retrieved from http://www.datalogforeningen.dk/fa/fa-20020221.html

Yuan, M.J. (2002, February 21). *Let the mobile games begin, Part 1. A comparison of the philosophies, approaches, and features of J2ME and the upcoming .NET CF.* Retrieved from http://www.javaworld.com/javaworld/jw-02-2003/jw-0221-wireless.html

Yuan, M.J. (2003, May 16). *Let the mobile games begin, part 2 J2ME and .NET Compact Framework in action.* Retrieved from http://www.javaworld.com/javaworld/jw-05-2003/jw-0516-wireless.html

KEY TERMS

Active Server Page (ASP): Microsoft's server-side technology that allows scripting language for dynamically generated Web.

Cell Phone (Mobile Phone): Electronic telecommunications device that is able to move over a wide area, connected using wireless radio wave transmission technology.

Personal Digital Assistant (PDA): Mobile device that serves as personal organizer. The basic features of a PDA include phone book, address book, task list, memo pad, clock, and calculator software.

Pocket PC: Operating platform for handheld devices introduced by Microsoft, based on the Windows CE operating system.

Smartphone: Mobile device that integrates the functionality of a mobile phone and PDA by adding telephone functions to a PDA or including the PDA capabilities on a mobile phone.

Web Service: Software system designed to support interoperability among machines over a network using a standardized interface.

Windows CE: A simplified version of the Windows operating system designed to run on handheld-size computers.

Windows Mobile: Operating system that replaced the Windows CE operating system on mobile devices. It includes a suite of basic applications for mobile devices based on the Microsoft Win32 API.

Standard–Based Wireless Mesh Networks

Mugen Peng
Beijing University of Posts & Telecommunications, China

Yingjie Wang
Beijing University of Posts & Telecommunications, China

Wenbo Wang
Beijing University of Posts & Telecommunications, China

INTRODUCTION

As various wireless networks evolve into the next-generation fixed broadband wireless access (BWA) systems, the wireless mesh network (WMN), expected as a promising technology, is still being standardized in IEEE 802.16 and commercialized in the World interoperability for Microwave Access (WiMAX) forum at present. In fixed BWA systems, the objective of applying mesh-typed topology is to build self-organized networks in the places where wired infrastructure is not pre-existing or not worthy to be deployed. The term "mesh-typed" here can also be described as "relay-based" or "multi-hopping," which means that the connection from a particular mesh subscriber station (mesh SS) to the mesh base station (mesh BS) is via one or more successive wireless links. Multi-hop wireless networking has traditionally led to significant research in the context of ad hoc or peer-to-peer networks. However, the fundamental goal of relaying augmented networks like WMNs is to provide wide-bandwidth coverage and high-data-rate throughput, while the defining goal of conventional ad hoc networks is to accomplish communications without any pre-existing infrastructure in a short time.

The mesh concept applying in WMAN systems has the relay-based and multi-hopping features. Since communications could take place through relay nodes, link distance could become much shorter, frequency and spatial reuse could become much more efficient, and interference could also become much lower. Thus WMNs could provide non-line-of-sight (NLOS) connectivity with high-data-rate capacity to extend the coverage range of existing point-to-multipoint (PMP) wireless networks, such as cellular mobile networks.

Figure 1 depicts a possible scenario where WMNs can be deployed to provide broadband access to the IPv6 backbone network. In WMNs, each mesh SS operates as not only a host but also a wireless router, which forwards transferring traffic within the network as well as traffic that goes out to other networks. The network is dynamically self-organizing and self-configuring, with both mesh BS and mesh SS automatically establishing and maintaining routes among themselves. All the nodes can use the distributed scheduling to ensure collision-free transmissions within their two-hop neighborhood, or use the centralized scheduling to complete functions in a more centralized manner through conveying much of the control work to the mesh BS; the combination of these two control mechanisms is termed as hybrid-controlling. Mesh BS is connected with WiMAX PMP BS through first tire wireless backhaul, and then WiMAX PMP BS is connected with the IPv6 backbone network through second tire wireless or wired backhaul.

In the centralized scheduling mode of WMNs, all traffic is restricted to be either in the direction of the mesh BS or away from the mesh BS. However, in the distributed scheduling mode, the transmissions are communicated between arbitrary pairs of nodes. Hence, an ad hoc network, described in Figure 1, could be considered as a type of uncoordinated distributed WMN.

Wireless sensor networks (WSNs) differ from the WMNs in that they contain hundreds or thousands of sensor nodes to allow for sensing over large geographical regions and these sensor nodes have much more limited computation capabilities, sensing capabilities, storage space, battery power, and transmission range. Even so, these sensors have the basic ability to communicate either among themselves or directly with an external BS, making WSN similar to both centralized and distributed WMN.

This article introduces a functional architecture supporting the wireless mesh networks for the IEEE 802.16 standard. Three essential techniques—collision avoiding, packet scheduling, and wireless routing—are intensively presented. Based on the mesh extension of the IEEE 802.16 medium access control (MAC) layer protocol and the relay-based characteristic of WMNs, the algorithms concerning those three essential techniques are briefly reviewed. The suitable algorithms for collision avoiding and packet scheduling mechanisms are analyzed. Meanwhile, the wireless routing algorithm for the proposal architecture is discussed. The future research work is presented and the research problems are focused.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. WMN internetworking architecture

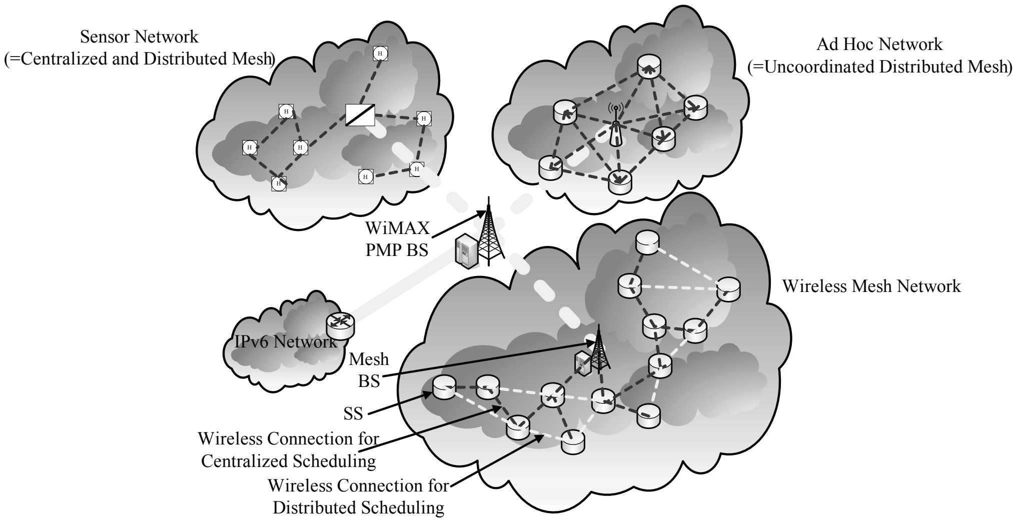
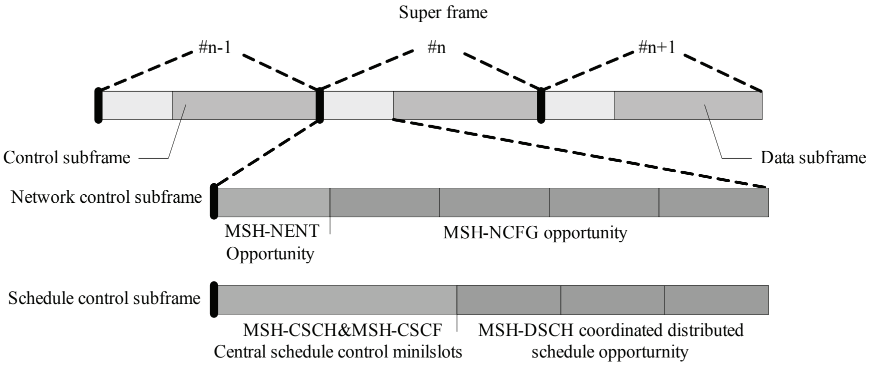

Sensor Network
(=Centralized and Distributed Mesh)
Ad Hoc Network
(=Uncoordinated Distributed Mesh)
WiMAX
PMP BS
Wireless Mesh Network
IPv6 Network
Mesh
BS
SS
Wireless Connection for
Centralized Scheduling
Wireless Connection for
Distributed Scheduling

Figure 2. Frame structure for IEEE 802.16 mesh mode


Super frame
#n-1
#n
#n+1
Control subframe
Data subframe
Network control subframe
MSH-NENT
Opportunity
MSH-NCFG opportunity
Schedule control subframe
MSH-CSCH&MSH-CSCF
Central schedule control minilslots
MSH-DSCH coordinated distributed
schedule opporturnity

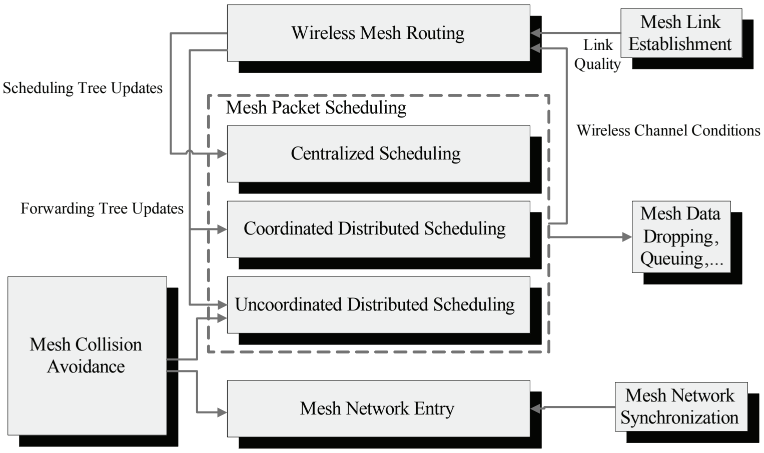
SYSTEM BLOCK

IEEE 802.16 mesh mode is an optional extension of the IEEE 802.16 MAC layer as an alternative or a complement to the conventional PMP architecture in the fixed BWA systems. The physical layer of it supports OFDM modulation, particularly as presented in the IEEE 802.16 standard with both the licensed frequencies and the license-exempt frequencies operating below 11G Hz. In this physical environment with long wavelength, requirements of line-of-sight (LOS) are not necessary and impacts of multi-path may be significant.

IEEE 802.16 mesh mode provides a novel method for new nodes to enter the network with the help of a full functionality member (sponsor node) in the network and has three kinds of scheduling modes including centralized scheduling, coordinated distributed scheduling, and uncoordinated distributed scheduling for efficient transmission of the data packets as well as control messages. The mesh frame structure defined in IEEE 802.16 mesh mode, which has no difference between uplink and downlink, is demonstrated in Figure 2.

We refer to the frame containing the network control subframe as the "network frame" and similarly term the frame including the schedule control subframe as the "schedule frame" for short. One network frame and several schedule frames constitute a super frame. In the network control subframe, the first opportunity with seven OFDM symbols is for NENT (network entry) message transmission in which a new node sends an entry request or entry acknowledgement. The symbols remaining are for NCFG (network configuration) message transmission in which the network configurations are advertised. In the schedule subframe every seven symbols

Figure 3. Functional architecture for IEEE 802.16 standard-based WMN



are grouped as a transmission opportunity. The first several transmission opportunities are utilized for central schedule messages, including CSCH (central scheduling) messages and CSCF (central scheduling configuration) messages, while the remains are for coordinated distributed schedule messages: DSCH (distributed scheduling) messages. DSCH messages may also appear in the data subframes during uncoordinated distributed scheduling. Data transmissions scheduled uncoordinatedly should submit to those scheduled coordinately.

FUNCTIONAL ARCHITECTURE AND RESEARCH AREAS

In order to focus on researching the key techniques of WMN, the suitable functional architecture, shown in Figure 3, is presented which is based on the cross-layer design and is to optimize the system performance. The proposed architecture consists of three main models: wireless mesh routing, mesh packet scheduling, and mesh network entry.

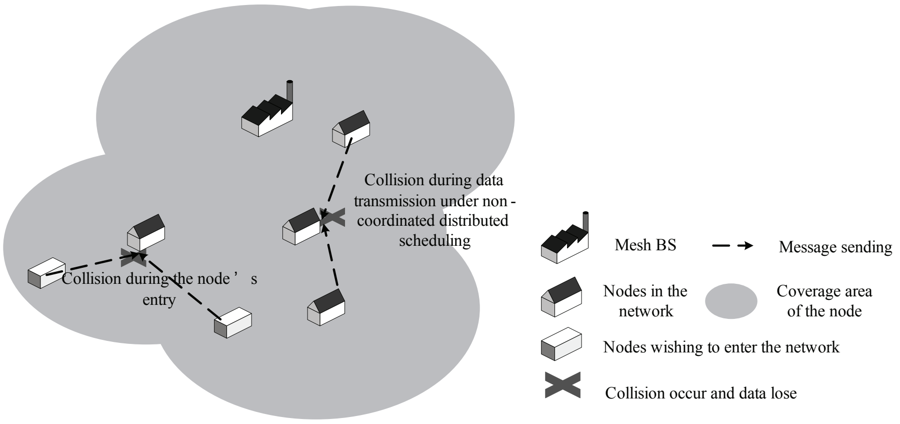
There are three sub-parts in the mesh packet scheduling model: centralized scheduling, coordinated distributed scheduling, and uncoordinated distributed scheduling, which are defined to represent these three scheduling mechanisms respectively. Meanwhile, in order to support the mesh routing procedure, the wireless mesh routing model is involved. Since both scheduling and routing algorithms require the wireless channel conditions to improve the throughput and meet the quality of service (QoS) requirement to support the various applications, there is a cross-layer design between wireless mesh routing and mesh packet scheduling models. The wireless mesh routing model will forward the routing message to the mesh packet scheduling and make the scheduling model update its scheduling tree periodically. When the radio channel condition is bad or there is not enough radio resource, the scheduling model can send a request message to the wireless mesh routing model to change the mesh routing. In order to guarantee the real-time requirement and minimize the delay, the directional and periodical signal exchange are necessary. Furthermore, the link measurement and establishment must be completed in the lower MAC layer, and the model mesh link establishment is added to assist the wireless mesh routing model.

What is more, the mesh collision avoidance model is proposed to avoid collisions that may occur during the mesh network entry process and the uncoordinated distributed scheduling process.

The mesh data dropping, queuing, etc., model is corresponding with the mesh scheduling mechanisms, while the mesh network synchronization model works together with the network entry procedure.

Among functional parts described in Figure 3, collision avoidance mechanisms, scheduling techniques, and routing algorithms are three challenging research areas. Most existing MAC protocols based on conventional collision avoidance mechanisms solve partial collision problems, but raise others. Considering particular features of wireless multi-hop mesh networks, how to improve existing mechanisms to avoid collision that would happen both in network entry procedure and in uncoordinated distributed scheduling is a crucial issue according to the scalability of the entire network. Since resources needed to be scheduled are dependent on transmission techniques applied by PHY layer, scheduling techniques must make adaptations correspondingly, especially when advanced techniques such as MIMO and cognitive radios are introduced. Routing in WMNs is also a tough task due to the delay-sensitive applications, as well as higher throughput and bandwidth requirements which distinguish these networks from other wireless multi-hop

Figure 4. Collisions in IEEE 802.16 mesh mode


Collision during data transmission under non-coordinated distributed scheduling
Collision during the node's entry
Mesh BS
Message sending
Nodes in the network
Coverage area of the node
Nodes wishing to enter the network
Collision occur and data lose

networks like ad hoc and sensor networks. Therefore, novel routing algorithms need to be proposed to utilize the potential advantages of the wireless medium.

KEY TECHNIQUES FOR PROTOCOL IMPLEMENTATION

Collision avoiding, scheduling, and routing can play a significant role in the implementation of IEEE 802.16 standard protocol. In this section we briefly review algorithms concerning those three key techniques and propose our new collision avoidance schemes, scheduling phases' division, and routing metrics exploration respectively.

Collision Avoidance Schemes

Although collision avoidance schemes have been presented in IEEE 802.16 mesh mode, collisions may still occur as demonstrated in Figure 4 during a new node entry process as well as during data transmissions controlled by uncoordinated distributed scheduling.

Solutions for alleviating collisions in the above situations could increase the network throughput and reduce the average delay to some extent. Since both the start time and the duration of a transmission are unpredictable, taking into account the low-cost-requirement of normal mesh devices, we focus our attention on random access schemes for WMNs rather than fixed resource allocation schemes or dynamic resource allocations-on-demand schemes.

Wireless Scheduling Mechanisms

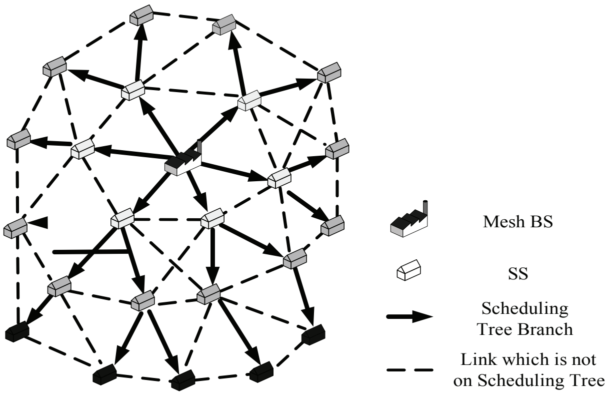
IEEE 802.16 mesh mode MAC supports three scheduling mechanisms: coordinated centralized scheduling, coordinated distributed scheduling, and uncoordinated distributed scheduling. Uncoordinated means performing in a partially, contention-based manner, while coordinated means using scheduling packets transmitted in a collision-free way within scheduling control subframes. Distributed means opportunity-based scheduling between two nodes, while centralized means mesh BS coordinates the radio resource allocation within the mesh network. In the centralized mechanism, every mesh SS sends its resource request to the mesh BS, and the mesh BS determines the resource allocation for each link and broadcasts the grants to SSs. A centralized mechanism is best for scheduling over links supporting persistent traffic streams, while distributed scheme is best for scheduling over links with occasional or brief traffic needs. Figure 5 depicts the scheduling tree for centralized scheduling and other possible connections for distributed scheduling. As shown in Figure 5, the centralized scheduled traffic only occurs on scheduling tree, in which different colors represent different hops from the mesh BS, while the distributed scheduled traffic occurs on both the scheduling tree and other links which are not scheduling tree branches. As most of the packet scheduling algorithms proposed for multi-hop wireless networks are TDMA based and could be further ameliorated for OFDM-based networks, our discussion will focus on the scheduling in a TDMA network.

The scheduling mechanisms discussed as follows are based on the following assumptions: nodes are assumed to use omni-directional antennas and operate in half-duplex mode, which means a node cannot transmit and receive in the same moment. The wireless channel is free of non-collision-related errors.

Wireless Routing Algorithms

In the case of IEEE 802.16 mesh mode, we consider routing in a relay-based network in which communication relations are limited to a few hops only. The choice of routing

Figure 5. Multiple scheduling mechanisms in IEEE 802.16 mesh mode



algorithm becomes challenging when considering multiple possible relays.

When an intermediate node in the proposed WMN architecture receives traffic, its routing function decides the next hop on the path to the destination, and this node forwards traffic along to the next hop node. In this type of relay-based network, traffic at a node waiting for routing decision may be on behalf of other nodes that are not within direct wireless transmission range of a mesh BS. This is the mechanism that is generally used for routing algorithm designing in ad hoc and sensor networks. Compared to ad hoc networks, WMNs applying relaying via fixed nodes do not need complicated distributed routing algorithms, while retaining the flexibility of being able to quickly select another route as a link between relays breaks. Based on the performance of the existing routing algorithms for wireless multi-hop networks and the specific requirements of WMNs, we examine the following issues to design an optimal routing algorithm for WMNs:

- **Routing Overhead:** The routing overhead in wireless mobile multi-hop networks may become severe because of substantial memory requirements and wide bandwidth spending for route establishing and maintaining, thus it results in low efficiency in network throughput. Studies of routing algorithms for mobile ad hoc networks (MANETs) by the MANET subgroup of the Internet Engineering Task Force (IETF) have shown high routing overhead and reduced potential efficiency of multi-hop techniques. However, for WMNs, limiting the number of hops in a range ($HR_{threshold}$, which is a configuration value that need only be known to the mesh BS, as it can be derived by the other nodes from the MSH-CSCF message) would greatly simplify routing complexity. The other factor that would also simplify routing complexity is nodes' minimal mobility. Taking into account these two factors, routing overhead would be reduced greatly. Based on these simplicity results from inherent characters of WMNs, the routing algorithm must minimize routing overhead as much as possible.
- **Fault Tolerance:** Some links between neighboring nodes may fail or be blocked due to physical damage or environmental interference. The failure of links should not affect the overall task of the WMNs to ensure robustness as one of the objectives to deploy WMNs. If a link breaks, the routing algorithm should be able to quickly select another route to avoid service disruption.
- **Scalability:** Since the number of user nodes in WMNs may be much larger than that in ad hoc networks, and routing in a relatively large wireless network must take convergence time and end-to-end delay into account, scalability is also an important routing design parameter in WMNs.
- **Load Balancing:** Relay-based is the most distinguishing characteristic of the WMNs. Thus the potential substantial network resources could be shared among many users through relay-based communications. In this cooperative way, load-sensitivity is an important factor to measure the relative performance of different routing metrics. For example, when a part of a WMN experiences congestion, new traffic flows on behalf of other users should be routed through other routes avoiding this part.
- **Routing Metrics:** Many existing routing algorithms for ad hoc networks have traditionally attempted to find routes by using shortest path as a routing metric. It is known that the defining goal of conventional ad hoc networks is to function without any pre-existing infrastructure, while the fundamental goal of mesh-typed multi-hop augmented networks is to provide wide-bandwidth coverage and high-data-rate throughput. Therefore, using shortest path metric for routing in multi-hop wireless networks is not sufficient to construct routes that are able to effectively transport traffic with reasonable delay, reliability, and throughput. In order to satisfy the original goal of WMNs, a routing algorithm must select better routes between given neighboring node pairs by explicitly using link quality measurements to explain the differences in quality of the paths. One obstacle to taking link quality into account is combining route metrics to form a path metric. This problem is not straightforward inherently, since several parameters might be used to give enough indication about the quality and behavior of a wireless link.

In order to exploit the potential advantages that the wireless medium defined by IEEE 802.16 offers, the approach employing cross-layer design is a good way to explore new routing metrics. The objective is to turn lower layer drawbacks into advantages through use of MAC and PHY information to help routing selection and making the routing layer control some lower layer settings.

SUMMARY AND FUTURE WEOK

The emergence of new multimedia and Internet applications for the wireless domain has spurred the study of WMNs for providing larger capacity and wider coverage. In this article, a novel functional architecture of WMNs based on IEEE 802.16 standard is proposed, and three key techniques demonstrated in this architecture, including collision avoiding, scheduling, and routing, are extensively studied.

While discussing the desirable features and suitable techniques, in addition to the basic properties of wireless relay-based networks, the special characteristics of IEEE 802.16 mesh mode, such as fixed and hybrid-controlling, are also taken into account. Since the new node entry process in IEEE 802.16 mesh mode involves no resource allocation or link establishment and there would be rare new nodes required to enter the network simultaneously. Considering the multi-hop topology and multi-service applications in this kind of network, the three-phase scheduling mechanism is presented and analyzed. Finally, we discuss the approach of cross-layer design for integrating the link quality into the routing metrics.

While we are designing the routing and scheduling algorithms, we should consider the physical limitations of each node and the interference between the transmissions of nodes in their neighborhoods. How to reflect the link condition from the physical layer to the network layer in time is a problem related to cross-layer design and needs more research work in the future. Besides that the path of a flow consists of multiple links, how to evaluate the overall performance of the route reasonably to choose the best routes is also an open issue. In addition, we need to consider more about the effective combination of scheduling and routing algorithms. Furthermore, there exist multi-services such as voice, minimum bandwidth required traffic flow, and best effort service. Therefore, QoS differentiation and guarantees must be supported.

Wireless mesh will transmit data packets based on all-IP protocol, and with the emergence of new services, the user of WMAN based on IPv6 will wish not only to access the Internet through wireless links but also to remain online even while they are moving. The WMAN technology based on IPv6 is therefore widely envisioned to have tremendous market potential. With this background, it would be meaningful to investigate the network protocol of IEEE 802.16 mesh mode and its integration with IPv6.

REFERENCES

Agis, E., Mitchel, H. et al. (2004). Global interoperable broadband wireless network: Extending WiMAX technology to mobility. *Intel Technology Journal, 8*(3), 173-187.

Beyer, D., Waes, N. V., & Eklund, K. (2002, February). Tutorial: 802.16 MAC Layer Mesh Extensions. *Proceedings of IEEE 802.16 Standard Group Discussions.*

Chen, J., Chi, C., & Guo, Q. (2005, October 3-5). A bandwidth allocation model with high concurrence rate in IEEE802.16 mesh mode. *Proceedings of the 2005 Asia-Pacific Conference on Communications* (pp. 750-754).

Erceg, V. et al. (1999). An empirically based path loss model for wireless channels in suburban environments. *IEEE/ACM Journal on Selected Areas in Communications, 17*(7), 1205-1211.

Erceg, V. et al. (2001). *Channel models for fixed wireless applications.* IEEE 802.16 Broadband Wireless Access Working Group, IEEE 802.16.3c-01/29r4. Retrieved from http://ieee802.org/16

Gupta, P., & Kumar, P. R. (2000). The capacity of wireless networks. *IEEE Transactions of Inf. Theory, 46*(2), 388-404.

Iannone, L., Khalili, R., Salamatian, K., & Fdida, S. (2004, September 20-22). Cross-layer routing in wireless mesh networks. *Proceedings of the 1st International Symposium on Wireless Communication Systems* (pp. 319-323).

IEEE Standard 802.16-2004. (2001). *Revision of IEEE standard 802.16-2001: IEEE standard for local and metropolitan area networks, part 16: air interface for fixed broadband wireless access systems.*

Ko, Y.-B., Shankarkumar, V., & Vaidya, N.H. (2000). Medium access control protocols using directional antennas in ad hoc networks. *Proceedings of IEEE INFOCOM 2000* (Vol. 1, pp. 13-21).

Li, J., Blake, C., De Couto, D., Lee, H. I., & Morris, R. (2001). Capacity of wireless ad hoc networks. *Proceedings of ACM SIGMOBILE.*

Nair, G., & Chou, J. (2004). IEEE 802.16 medium access control and service provisioning. *Intel Technology Journal, 8*(3), 213-228.

Pabst, R. et al. (2004). Relay-based deployment concepts for wireless and mobile broadband radio. *IEEE Communications Magazine, 42*(9), 80-89.

Sun, Z., Lu, Y., Zhou, Y., Peng, M., & Wang, W. (2005). A simulation model for the IEEE 802.16 broadband wireless access systems. *OPNETWORK,* (8).

Sun, Z., Lu, Y., Zhou, Y., Peng, M., & Wang, W. (2005). Research of uplink packet scheduling mechanisms based on GPSS in IEEE 802.16 systems. *Proceedings of ICCI.*

Wu, Z., Liu, M., Wang, Y., Li, M., Peng, M., & Wang, W. (2005). Investigation of collision avoidance mechanisms in the IEEE 802.16 based wireless mesh networks. *OPNETWORK,* (8).

KEY TERMS

CSCF: Central scheduling configuration message.

CSCH: Central scheduling message.

DSCH: Distributed scheduling message.

$HR_{threshold}$: A configuration value that need only be known to the mesh BS, as it can be derived by the other nodes from the CSCF message.

NCFG: Network configuration message.

NENT: Network entry message.

Network Frame: The frame containing the network control subframe to "network frame."

Schedule Frame: The frame containing the schedule control subframe to "schedule frame."

Super Frame: One network frame and several schedule frames constitute a super frame.

Taxonomies, Applications, and Trends of Mobile Games

Eui Jun Jeong
Michigan State University, USA

Dan J. Kim
University of Houston Clear Lake, USA

INTRODUCTION

Wireless communications and the distribution of cell phones have been rapidly extended with the expansion of mobile content services since the early 2000s. With such extension, mobile games have been viewed as a separate branch in game device platforms. While studies on mobile contents have increased for several years, research on mobile games is still in the early stages. Although mobile games have developed and expanded their ranges in game markets, there is little research on the classification and development trend of mobile games. Considering that game devices have been converged into ubiquitous communication/networking features and the range of games has been expanded from entertainment to education, health, and exercise, there is an urgent need to study mobile games' taxonomy, application, and future trends.

In this article, mobile games are classified by several criteria (i.e., contents, platforms, and multi-layer based). Examples of mobile games are summarized, along with taxonomies. Lastly, applications and the macro trend of mobile games will be presented. In addition, some insights in the design and development of mobile games will be discussed.

MOBILE GAME TAXONOMIES

Games are different from other genres such as music, film, and literature in the participation of users. *Interactivity* and *narratives* are two important factors to categorize games. Aarseth, Smedtad, and Sunnana (2003) classified games with a number of basic dimensions such as space, perspective, time, and teleology. Klabbers (2003) suggested social systems such as actors, rules, and resources for the establishment of game taxonomy. Wolf (2005) considered some standards such as the games' goals and objectives, and the nature of the games' play-characters and control devices. The devices have been applicable to traditional games in PC or console games with wide monitors, gorgeous graphic environments, and broad structures of narratives. Nowadays, however, with the development of fusion games and the acceleration of genre convergence, the clear division of game genres has been difficult because many cross-listed games emerged in two or more genres (Wolf, 2005). Considering such a trend and characteristics of mobile games, this article classifies mobile games into some basic genres.

First, content-based taxonomy is conducted from the basic features of games such as the control range of gamers, the role of characters, and the degree of user participation. Second, platform-based is from mobile device platforms with which games are played. Third, multi-layer-based is from the capability of multi-player network and 3D graphic technology.

Content-Based Taxonomy

The role of gamers is the essential element in game taxonomy. Gamers can take their own individual roles or become omnipresent beings in games. Role-playing games (RPGs) comprise the representative genre of individual role games; strategic simulation games are controlled by omnipresent gamers. These games can be divided by the environments of the role of gamers: gamers should be a shooting gunner in shooting games; gamers should be an adventurer in an unknown world in adventure games. In these games, users usually take on their own roles. Finally, such games could be classified by the degree of user participation: multi-player games are played by the collaboration of individual roles; team games are played by teams (or guilds) with enough members having individual roles.

- **Role Playing Games:** A gamer as a character takes on an individual role in accomplishing missions or quests. The user can upgrade his/her character and take items to complete missions effectively.
- **Simulation Games:** Users complete their missions in simulated environments by controlling resources such as objects, characters, and items with their own strategies. There are some types of simulation games such as construction, management, or war simulation games.
- **Fighting Games:** Gamers take a character and fight with skills of kicking or striking against other char-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

acters to win the contests. There are some kinds of fighting games mixed with action marshals such as judo and boxing; they are sometimes referred to as action fighting games.

- **Shooting Games:** Users take on their missions as a shooter or artilleryman in a war or as an infiltrating agent in an operation. In these games, the perspective of a user is a very essential element in attracting the user's involvement. Therefore, shooting games with the first-person perspective are usually referred to independently as first-person shooting games (FPSs).
- **Adventure Games:** Users take travels to unknown space or environment as travelers or warriors.
- **Sports Games:** Users take on a role or control teams in sports contests such as baseball, basketball, or football. Racing sports such as riding and car racing are usually called racing games.
- **Board Games:** Users compete with opponents in traditional board games such as chess, Tetris, puzzles, oriental chess, baduk, and so forth.
- **Single-User Games:** Only one user can participate with a role or mission.
- **Team (or Guild) Games:** Users should join a team with other users to complete missions or win a contest.
- **Massively Multi-Player Online Games (MMOGs):** A huge number of users can participate simultaneously with their roles or missions.

Platform-Based Taxonomy

Mobile devices are also regarded as independent platforms. Each mobile device has its own features in containing mobile games. Thus, if producers want to transplant a game in a device into another one, they should restart the product processes from the beginning. For this reason, platform-based taxonomy is beneficially used in mobile games. In terms of mobile devices, mobile games are classified into mobile phone, portable console, and PDA games. In terms of mobile platforms, mobile games are called Java games, Brew games, and WAP games.

- **Mobile Phone Games:** Conducted in cell phones.
- **Portable Console Games:** Played in portable consoles; examples include PSP (Play Station Portable), NDS (Nintendo Dual Screen), and GBA (Game Boy Advance).
- **PDA Games:** Embedded or downloaded in PDAs.

Multi-Layer-Based Taxonomy

Multi-layer-based games are classified with the adaptation of high technology including capability of multi-player network and 3D graphic technology. Mobile games have developed from text-based to 3D multi-user network games. Therefore, in terms of multi-layer-based taxonomy, mobile games can be classified into five types: text-based, 2D graphic, 2D network, 3D half network, and 3D multi-user network games. Table 1 summarizes the taxonomies of mobile games with genre examples by the criteria of division.

MOBILE GAME APPLICATIONS AND INDUSTRY

Application Areas of Mobile Games

The application areas of mobile games can be categorized into four areas: traditional game industry, mobile Internet applications, mobile advertisements, and new applicable areas.

The most applicable area of mobile games concerns the traditional game industry. With the development of handheld

Table 1. Taxonomies of mobile games

	Criteria of Division	Genre Examples	Examples
Content-Based	Control range of gamers	RPGs Simulations	*Doom RPG* *Real Estate Tycoon*
	Role of characters	Fighting (Action) Shooting, Sports Adventure, Board	*Mortal Kombat* *Quake Mobile, FIFA06* *Tomb Raider, Tetris*
	Degree of user participation	Single-user, Team, Multi-user	*Deep Pocket Chess* *Samgukji, Undercover2*
Platform-Based	Device platforms	Cell phones, PDAs, Portable consoles, etc.	
	Game platforms	WAP, Java, Brew, etc.	
Multi-Layer-Based	Adaptation of 3D graphic and network technology	Text-based, 2D graphic, 2D network, 3D half network (3D games with two or several users) 3D multi-user network (3D MMOGs, etc.)	

mobile services, PC and console games have transplanted popular games into mobile devices: old big hit games have found their new hit area in game markets, so big game companies have expanded their business areas into mobile games. Owing to the advanced technology, some traditional game devices have included mobile capabilities. The boundary between mobile phone games and portable console devices has been revolutionarily converged, and mobile game cartridges or multimedia cards (MMCs) have been sold in game shops with game CDs and DVDs. Mobile games have provided various channels of game services. Internet portals and interactive TV service companies have provided mobile game services, and console game manufactures have opened mobile portal services for their customers of portable console devices with network capability. In particular, game producers developed new games such as LBS (location-based service) games using the features of mobile devices, and old mobile games have been upgraded to multi-user 3D games with the development of technology. Most of all, mobile games have been used by game companies to provide further services to users. Gamers can use their items and avatars saved in PC or console games by using mobile games, because PC and console network games can be used in mobile devices at any place, and they are linked to original devices without any loss of user information.

Secondly, mobile games have great potential making mobile Internet applications flourish. Mobile game communities could expand mobile communities: the conversion of Internet games into mobile games could drive many Internet game communities to change their main space into mobile communities. Like the online shopping malls connected with Internet game portals, mobile game users could be excellent resources for shopping malls in mobile services; as with Internet chatting rooms in game portals, mobile chatting services could be developed with the connection of mobile games. With the division of mobile game users into heavy and light users, mobile portal sites would be differentiated: mobile game user types could be applied to analyze mobile user types for business, because game users could account for a high ratio of heavy mobile users. For the security of big hit mobile game information, mobile security technology and anti-duplication technology of mobile content sources would be applied into mobile games. Mobile game events such as mobile game exhibition, mobile game contests, and mobile game character shows could expand areas of mobile Internet applications.

Thirdly, as mobile game users are not restricted in terms of age, gender, and social status, advertisements through mobile games could be as effective as public broadcasts. Companies can inform consumers of their logos or specific brands through mobile games. Recently, public relationships using the characters of popular games have been attracting the attention of companies, and information services about new products for target customers have been provided through SMS (short messaging services), EMS (enhanced messaging services), or MMS (multimedia messaging services).

Finally, mobile games could be applied to new areas such as education, exercise, and therapy in mobile services. Games could be a good way to enhance participation and involvement, while mobile devices could provide both easy accessibility and convenience. So, mobile services in the form of interactive games such as game-applied education, exercise programs with interactive games, and health programs with MMS could be developed.

Mobile Game Industry

Mobile games comprise a rapidly developing industry in the game markets. In the market of mobile phone games, according to Datamonitor (2002) and In-stat/MDR (2004), the market size estimated around $500 million in 2002 is expected to grow over $5 billion in 2008. The size of the U.S. market was about $40 million in 2002, but it is expected to exceed $1 billion in 2008. Mobile game users are expected to reach over 70 million in 2008, which is about 10 times higher than in 2002. These figures exceed the growth ratio of other game markets such as PC-online and console games. As the economy of developing countries improves, and mobile devices continue to spread, mobile games are increasingly regarded to comprise a most fascinating market. The Asia-Pacific market has captured over 50% of the total mobile phone game market, with Japan and Korea in the lead, but China, India, and South Asian countries have been expanding their market share at a speed parallel to the spread of mobile devices. In portable console games, according to NPD (2006), revenue in 2005 was $1.4 billion, notwithstanding the stagnation of console game markets in the U.S. Portable console markets have grown rapidly both in Europe and Asia with the development of network games.

Mobile games have been developed with the spread of mobile devices, gradual upgrade of capabilities of mobile devices, development of graphic technology, introduction of the flat sum system, and the spread of mobile cultures. The mobile game market is no longer independent from other game markets, because most traditional games in PCs and consoles are translated into mobile devices, and games are converged without the division of game machines. The advent of game-specialized mobile phones and portable console game devices with network capability has accelerated the convergence of games. World game publishers have merged mobile game aggregators and extended the ratio of mobile game products. Leading game machine manufacturers (platform holders) such as MS, Sony, and Nintendo are competing for dominance of the future market, not only of game devices, but also of home entertainment devices. They are striving to take the world standard of new media with their game machines. Mobile game devices will be the ultimate destination for business triumph, because mobile devices are the most prevalent tools used by game users, and mobile games would be the most useful and profitable content in the entertainment market using state-of-the-art technology.

FUTURE TRENDS AND CONCLUSION

Several trends in the mobile games industry are expected in the near future: the convergence of game devices, variation in game content, and prevalence of mobile game cultures. The first outstanding change will be the acceleration of device convergences into mobile game devices. As console game machines have been developed into portable devices, mobile phones will be enhanced with high-capability games, and new mobile game devices based on PC capabilities will be created. To conquer preoccupation of future mobile device markets, game publishers will focus on popular game brands and develop new technologies with high multimedia functions. Game producers will create less-expensive 3D network games and more various genre games. User interfaces in mobile devices will be enhanced, and devices will be focused on game functions that can massively run multi-user games. Long-lasting charge cells will be developed for mobile devices, and local area network (LAN) games with Wi-Fi and Bluetooth will be pervaded. High-definition screens will be able to run previous PC or console games with multimedia cards and security digital (SD) cards.

The second trend is variation in game genres. Games will be differentiated for heavy and light users: for heavy users, hit MMORPG or strategic games such as *World of Warcraft, Starcraft,* and *Lineage* series will be created as mobile games. Mobile portals will be differentiated in terms of specified services, and their services will include generic mobile services such as shopping, chatting, and transactions. With the diffusion of flat sum services and the stabilization of the mobile network, 3D network games could be used in both mobile devices and PC or console devices simultaneously.

The third trend is the expansion of the mobile game cultures. As game contests and exhibitions are popularized in game markets, mobile game events would gain more space among them. Mobile games will extend their ranges and lead mobile content services, and they will be regarded as one of the best ways to provide mobile interactive services such as education, medical consultation, and exercises. Therefore, new mobile games using such mobile cultures will be created beyond entertainment.

In short, mobile games expand their areas not only with the convergence of game devices, but also with the expansion of their areas of application. With the development of state-of-the-art mobile devices and network technology, mobile games can include all the games played in other device platforms such as PCs, consoles, and arcade game machines with the same environment. MMOGs in online computer games and high-definition 3D games in consoles can be adapted into mobile games. Furthermore, with the prevalence of mobile cultures, mobile games are thought of as a tool to communicate with others in various ways. As mobile cultures such as mobile communities expand, mobile games have expanded their ranges from entertainment to more applicable services such as education, exercise, and industry advertisements. Mobile games will expand their shares not only in game markets, but also in mobile-applied content markets.

REFERENCES

Aarseth, E., Smedtad, S. M., & Sunnana, S. (2003). Multi-dimensional typology of games. *Proceedings of the Level Up Conference* (pp. 48-53). Utrecht: University of Utrecht.

Brad, K., & Borland, J. (2003). *The rise of computer game culture: Dungeons and dreamers from geek to chic.* New York: McGraw-Hill.

CESA. (2005). *2005 CESA game white paper.* Tokyo: Computer Entertainment Suppliers' Association.

Datamonitor. (2002). Asia-Pacific mobile gaming: A study of best practice. Identifying success factors the Asia-Pacific markets. *Datamonitor,* (October).

Gee, J. P. (2003). *What video games have to teach us about learning and literacy.* New York: Palgrave Macmillan.

Hall, J. (2005). Future of games: Mobile gaming. In J. Raessens & J. Goldstein (Eds.), *Handbook of computer game studies* (pp. 47-55). Cambridge, MA: MIT Press.

In-Stat/MDR. (2004). Mobile gaming services in the U.S., 2004-2009. *In-Stat/MDR,* (August).

KGDI. (2005). *2005 game white paper.* Seoul: Korea Game Development & Promotion Institute.

Klabbers, H. G. (2003). The gaming landscape: Taxonomy for classifying games and simulations. *Proceedings of the Level Up Conference* (pp. 54-67). Utrecht: University of Utrecht.

Newman, J. (2004). *Videogames.* London: Routledge.

Nokia. (2003). *Introduction to mobile game development.* Retrieved from www.forum.nokia.com/html_reader/main/1,,2768,00.html

NPD. (2006). *The NPD group reports annual 2005 U.S. video game industry retail sales.* Retrieved January 17, 2006, from www.npd.com

Ring, L. (2004). *The mobile connection: The cell phone's impact on society.* San Francisco: Morgan Kaufmann.

Schwabe, G., & Goth, C. (2005). Mobile learning with a mobile game: Design and motivational effects. *Journal of Computer Assisted Learning, 21,* 204-216.

Wolf, M. (2005). Genre and the video game. In J. Raessens & J. Goldstein (Eds.), *Handbook of computer game studies* (pp. 193-204). Cambridge, MA: MIT Press.

KEY TERMS

Game Genre: One of several classified game types in terms of storylines, role of players, or device platforms.

Game Publisher: Game provider and copyright owner who connects game producers and content providers.

Local Area Network (LAN) Game: A network game played with other users within a short distance using LAN equipment such as Wi-Fi and Bluetooth.

Multi-Layer Game: A multi-user game where high-dimensional graphic and network technologies are adapted, such as a 3D full-network game.

Platform Holder: A company that provides game platforms to developers for development of games. In console games, device manufacturers are the same as platform holders.

WAP Game: Game serviced by wireless application protocol (WAP).

A Technology Intervention Perspective of Mobile Marketing

Dennis Lee
The University of Queensland, Australia and
The Australasian CRC for Interaction Design, Australia

Ralf Muhlberger
The University of Queensland, Australia and
The Australasian CRC for Interaction Design, Australia

INTRODUCTION

In the last decade, the explosive growth and adoption of mobile phones has become commonplace in our everyday lives (Haghirian, Madlberger, & Tanuskova, 2005). In 1997, there were only 215 million people worldwide who used mobile phones as communication devices (Bauer, Barnes, Reichardt, & Neumann, 2005). Today, it is estimated that 2 billion people own a mobile phone worldwide and this number makes up a third of the entire human population (Wireless Intelligence, 2005).

Mobile phones are no longer thought of as mere personal communication tools (Cheong & Park, 2005; Ito & Okabe, 2005). They have become a fashion symbol for teenagers and young adults (Katz & Sugiyama, 2005). Personalised ring tones, colours, display logos and accessories are individualised accordingly to suit individuals' preferences (Bauer, Barnes, Reichardt, & Neumann, 2005). Furthermore, mobile phones are no longer just a platform for voice calls and sending and receiving text messages such as short messaging service (SMS). Photos, pictures and video clips can be attached as a multimedia message service (MMS) for communication purposes too (Okazaki, 2005a). With the recent introduction of 3G mobile technology, mobile phone users are able to perform more activities via their 3G enabled phone sets. They are able to browse the Internet fairly quickly, access online banking, play video games wirelessly, watch television programs, check for weather forecasts, allow instant messaging, and perform live video-conferencing (Okazaki, 2005b).

The rapid growth of the mobile industry has created a foundation for mobile commerce (m-commerce). M-commerce facilitates electronic commerce via the use of mobile devices to communicate and conduct transactions through public and private networks (Balasubramanian, Peterson, & Jarvenpaa, 2002). The current emerging set of applications and services that m-commerce offers include mobile financial applications, mobile entertainment and services, product locating and shopping, wireless engineering, mobile auctions, wireless data centres and mobile advertising (Malloy, Varshney, & Snow, 2002). Commercial research has indicated that consumers' interest in m-commerce services and mobile payments have increased from 23% in 2001 to 39% in 2003 (Harris, Rettie, & Cheung, 2005). It is projected that by 2009 the global mobile commerce market will be worth at least US$40 billion (Juniper Research, 2004).

Considering the projected worth of mobile commerce and the number of mobile subscribers, mobile marketing is increasingly attractive, as companies can now directly convey their marketing efforts to reach their consumers without time or location barriers (Barnes, 2002). The potential of using the mobile medium to market is now more attractive than before (Karjaluoto, 2005), as it can assist companies in building stronger relationships with consumers (Barwise & Strong, 2002), and can be used as a promotional channel to reach consumers directly (Barnes, 2002; Kavassalis, Spyropoulou, Drossos, Mitrokostas, Gikas, & Hatzistamatiou, 2003; Okazaki, 2004) anywhere and anytime.

However, many aspects of mobile marketing are still in its infancy (Bauer, Barnes, Reichardt, & Neumann, 2005; Haghirian, Madlberger, & Tanuskova, 2005; Okazaki, 2004, 2005b; Tsang, Ho, & Liang, 2004). Research into mobile marketing is currently lacking, as this is a relatively new phenomenon. Very few studies have been conducted to demonstrate how the mobile phone channel can be successfully integrated into marketing activities of companies (Balasubramanian, Peterson, & Jarvenpaa, 2002; Haghirian, Madlberger, & Tanuskova 2005). Furthermore, no studies to date have compared the effectiveness of this mobile medium in delivering advertising and sales promotion with other more established media such as the print medium.

The fundamental question that remains unresolved is, "What is the difference between mobile marketing and traditional marketing?" Will this new form of marketing be effective? How will consumers respond to this form of marketing? What will be the benefit to marketers when consumers receive this type of advertising? These are just some of the issues that marketers are concerned with in order to

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

evaluate the mobile channels for marketing purposes and are questions that are core to computer-supported collaborative work (CSCW) and technology intervention research.

MOBILE MARKETING

Mobile marketing via SMS-based advertising and sales promotions is now being carried out by several multinational corporations (MNCs) in Europe and the United States of America. MNCs are very cautious in integrating such a new medium into their marketing mix (Mayor, 2005; Okazaki, 2005a). This is mainly because marketers are not fully convinced of the value of mobile channels as a marketing tool (Haghirian, Madlberger, & Tanuskova, 2005). Marketers are unsure whether their marketing efforts will cause positive or negative impacts on their consumers.

Another issue is the difference in worldwide telecommunication networks and mobile handsets used in the last decade (Leppaniemi & Karjaluoto, 2005). The recent introduction of 3G mobile technology as a worldwide standard for telecommunication networks and mobile handsets has brought about a new level of investment safety for companies (Karjaluoto, 2005). Companies are beginning to test their marketing efforts via the mobile phone medium (Cheong & Park, 2005). This suggests the need for researchers to develop theories and models to inform how mobile marketing can work effectively in the mobile phone context (Karjaluoto, 2005).

According to Tsang, Ho, and Liang (2004), mobile marketing can be classified as either permission-based, incentive-based or location-based. Permission-based marketing requires mobile users' prior approval before specific marketing messages can be sent (Barwise & Strong, 2003). By getting the permission of the mobile users, the factor of irritation may be reduced when users read the advertisement. Incentive-based marketing provides specific rewards to individuals who agree to receive promotions (Tsang, Ho, & Liang, 2004). For instance, mobile phone users may get free connection time from their mobile service providers for retrieving and reading advertisements. Location-based marketing targets mobile users in a certain location. The advantage of location-based marketing is that advertisements are sent to those individuals who are present or near the location (Barnes, 2003).

The incentive-based marketing approach is adopted because most consumers perceive the current mobile marketing as advertising, without making a distinction between sales promotions and advertising messages (Gogus, 2004). In other words, most consumers will generally term any marketing message received on their mobile phones as an advertisement, regardless of content (Gogus 2004). Moreover, the dominant form of mobile "advertising" appears to be in the form of promotion (Kavassalis, Spyropoulou, Drossos, Mitrokostas, Gikas, & Hatzistamatiou, 2003; Haghirian, Madlberger, & Tanuskova, 2005; Mayor, 2005; Okazaki, 2004; Tsang, Ho, & Liang, 2004).

In the marketing literature, a sales promotion can be defined as a more direct form of persuasion that may offer incentives to stimulate immediate purchase behaviour (Rossiter & Percy, 1998). Examples of sales promotional incentives include coupons, on-pack promotions, bonus packs, samples, premiums, and sweepstakes (Rossiter & Bellman, 2005; Shimp, 2003). Most of these promotional tools are based in print and termed as traditional promotional incentives (Belch & Belch, 2004).

On the other hand, advertising can be defined as a relatively indirect form of persuasion that may cause a favourable mental impression and then create an inducement toward a purchase response (Rossiter & Percy, 1998). Advertising is considered as the placement of a message to either increase product awareness, promote sales of goods and services, or just disseminate information (Leppaniemi & Karjaluoto, 2005). Advertisements may also include the element of sales promotion, a common example of which is in the form of coupons.

Coupons are considered to be some types of inducement that provide extra incentives to buy (Belch & Belch, 2004). Thus, in the context of mobile promotion, a mobile coupon is defined as an incentive that is paperless and electronic in nature (Wehmeyer & Müller-Lankenau, 2005). It is the fusion of the traditional print-based coupon with the mobile phone medium. A mobile coupon is delivered to a mobile phone handset as a message and is associated with mobile services and contents (Wehmeyer & Müller-Lankenau, 2005).

INTERACTION DESIGN AND THE LOCALES FRAMEWORK

The Locales Framework is a comprehensive theoretical CSCW and interaction design framework in the field of information and computer science (Fitzpatrick, 2003). According to Fitzpatrick, Kaplan, & Mansfield (1998), this research framework is an approach that allows for the creation of shared abstractions among stakeholders (e.g., companies, individuals, consumers, marketers), and also to narrow the gap between social and computing concerns with a common language. Understanding the social phenomenon and designing a relevant application that can fit the social setting are the two important factors when applying the Locales Framework. It is the aim of Locales Framework analysis that more pragmatic design and systems applications are built to suit the social world (Fitzpatrick, 2003).

The Locales Framework is based on five aspects, each of which are interdependent and overlapping, as they share various concerns with one another and are used to approach the domain to be studied from different perspectives—rather than separating the domain into distinct subdomains to be studied independently.

The *locale foundation* aspect portrays the social world and the locale it uses for its interaction (Fitzpatrick, 1998, p. 91). The social world can be characterised by a number of issues such as collective goal, memberships, duration, structure, culture and roles. A locale is the primary unit of analysis in the Locales Framework. A locale consists of the site and means that a social world uses in its pursuit of the shared purposes. According to Fitzpatrick (2003), a site is a place the social world uses and means are the objects within this place. The social world needs sites and means to facilitate their shared interactions.

The *civic structure* aspect takes the locale of interest and considers its relationships and interactions with the wider community (Fitzpatrick, 1998, p. 92). In other words, it concerns the facilitation of interaction with the wider community within and beyond a person's known social worlds and locales. The interaction with a wider community can possibly relate to an environment that is physical, spatial, geographical, organizational, informational, professional, legislative, and so on.

The *individual view* aspect describes an individual's single perspectives on one social world as well as on multiple social worlds (Fitzpatrick, 1998, p. 115). A single perspective is how an individual sees one social world, and is dependent on the level of engagement with the centre of that world, whereas multiple view sets incorporate the individual's views of all the social worlds with which he or she is engaged. Individuals personalize their views to suit their tasks according to their current level of engagement.

The *interaction trajectory* aspect identifies the dynamic, temporal aspects of the social world in action (Fitzpatrick, 1998, p. 122). This aspect identifies the actual interactions individuals have over time within the setting and with each other. Moreover, this aspect is not only concerned with the current action, but also with the past and projected futures. Awareness of past actions and outcomes, present situations, and visions for the future are important for creating plans and strategies. An important consideration to understand this aspect is to look at what perspective or point of view is applied to any particular domain.

The *mutuality* aspect is a collaborative activity that draws specific attention to how the locale supports presence, and how awareness of that presence is supported for the achievement of shared activity. The mutuality aspect enables questions on who, what, when, where, why and how to be answered.

When the Locales Framework is used, it involves a two-phase approach. This is iterative in order to better understand the nature of the given (Fitzpatrick, 1998). The first phase is to understand the current locales of interest from the view of the interaction needs. This could involve using qualitative data collected through an ethnographic study or a one-to-one interview. Generally the data collected could then help to provide some relevant structure to designers when they engage in the design process of an application. It is argued that for designers who do not have any social science background, the Locales Framework could be applied as a sensitizing device to aid in formulating initial questions for the design process (Fitzpatrick, 1998).

The second phase in applying this framework is to evolve new locales. The goal is to discover more possibilities for the existing locale of interest in order to better support the activities that take place there and to explore possible newer locales that can evolve as a result. This phase is to identify the advantages of any available medium, physical or computational, and the synergy among them, so that the needs of the social world are better meet. Specific questions that will help to drive this phase include: What interaction needs does the social world need that are lacking in this current locale? How can the existing locale be enhanced to support the aspects of the Locales Framework; namely, mutuality, individual views, civic structure and interaction trajectory? Can new technology be applied to the locale? Can new social worlds evolve if the resources are used in newer locales?

MOBILE MARKETING AS TECHNOLOGY INVENTION

Prior research has identified the importance of coupons in affecting consumers' cognitive, affective and conative behaviour during promotional campaigns (Raghubir, Inman, & Grande, 2004), but relatively little research has been conducted into the use of electronic coupons (Fortin, 2000; Suri & Swaminathan, 2004), particularly the form of mobile coupons (Okazaki, 2004). Most research in coupon studies is based mainly in the traditional medium of print (Coyle & Thorson, 2001; Liu & Shrum, 2002).

Much of the current literature that has been mentioned is adapted purely from a marketing perspective. Since mobile marketing involves people, technology and applications, mobile marketing should also be investigated from a human-computer interaction (HCI) and CSCW perspective. This will perhaps provide a better understanding of how and what is best for mobile marketing.

To better design mobile marketing strategies from a technological viewpoint, the use of the Locales Framework can be applied. The five inter-dependent characteristics of the Locales Framework guide study of the product or service to be marketed. An example may be a coffee shop:

Locale Foundation

The social world will be portrayed by consumers trying to buy beverages or food at the cafes and the locale is the cafe. The means in this case will be the chairs, tables, coffee machines, coffee counters, and the cashier's machine found within this site. The new technologies, that is, mobile phones and systems to send mobile coupons, are also included. More broadly the cafe may be situated in a locale such as a shopping centre or University that has its own means.

Civic Structure

The civic structure aspect considers issues such as the physical location of the café in its broader situation, store layout, and any competitors of café.

Individual View

In the case of mobile marketing, the individual that comes into the café will be a consumer and thus his or her task may be to purchase a cup of beverage for enjoyment. When they leave the café, their perspective may change to follow the priority that they may have to engage in. Perhaps the perspective may change to acquire knowledge and thus attend lesson at a lecture theatre or maybe need to catch a ride home by become a passenger when boarding a public transport such as a bus.

Interaction Trajectory

In this case study, the interaction trajectory aspect will determine the objective of the consumer coming to the café and how does the consumer interact with the surrounding environment. Despite receiving a discount coupon for cheaper beverages via the mobile phone, the consumer may come into the café with the purpose of meeting someone and not buy any beverages at all. To this consumer, the café has become a meeting venue and not a place for consumption. The café may become a place for taking a coffee break with fellow colleagues, and therefore the consumer may take advantage to purchase a cup of coffee at the special price for enjoyment.

Mutuality

In this case scenario, the mutuality factor will look into how mobile marketing is supported and how applications can be created to support mobile marketing in the context of a café.

Several advantages of the Locales Framework, as according to Fitzpatrick (1998, pp. 152,153), are listed as follow:

- It provides a common tool for understanding and designing of a social problem.
- It has the potential to analyse issues from group to individual level, local setting to global context, and structure to process;
- It is independent of any one theoretical orientation when investigating into one phenomenon.
- It is a framework that is strong in identifying key elements of a collaborative environment but sufficiently generic, open and incomplete so as not to prescribe nor circumscribe all that is of interest.

Applying the Locales Framework approach to mobile marketing, we are able to build several "locales" (which can be defined as potential social world scenarios) in helping companies to consider before they actually implement their marketing plan. Moreover, companies that implement mobile marketing should consider the aspect of civic structure—the facilitation of interaction among various factors like physical, informative, geographical and technological parties. In the context of mobile marketing, companies should look at who their telecommunication service providers are, where their potential consumers are located, what applications should be used to generate response from consumers and what types of mobile phones should be able to receive mobile marketing. Companies need to understand that their potential consumers have many different perspectives and opinions, another aspect considered in the Locales Framework analysis.

As mentioned earlier, the inducement of using a coupon to induce potential consumers to respond to mobile advertisements is a possible suggestion. Furthermore, companies need to understand that the locale does not stay static, as it is always changing and evolving. Therefore companies need to involve the interaction trajectory aspect that identifies actual interactions individuals have over time within a given context setting and with each other. The last aspect on mutuality involves companies to consider how best mobile marketing can be supported in a given location and how mobile marketing can create awareness for the companies.

The Locales Framework does not attempt to account the findings of one particular phenomenon that is generalisable. It does, however, deliberately aim to characterise its findings that are open in many ways. In fact, one of its aims is to focus on providing an evolvable framework that can be made relevant for both understanding the social situation (in this case mobile marketing) and for improving better technology development to it. This approach, unlike many traditional marketing strategic research approaches, does not assume that the technology introduction has to accept static technologies, or unchanging user attitudes towards technology (rather than the product). A dynamic, multi-dimensional picture of clients and possible interactions allows more dynamic engagement models—supported by dynamic technology, not based on working around systems controlled by other developments.

FUTURE TRENDS

Mobile marketing still lacks research. However, approaching this area from a technological perspective we can suggest several possible outcomes that a marketing and CSCW combined approach indicate:

First, companies are increasingly able to understand potential consumers' attitudes and behaviours in the context of mobile marketing from a more holistic perspective. In particular, the adoption of Locales Framework can provide insights to companies on how to further improve their design and concept for mobile marketing strategies in a particular situation.

Second, there is a need to consider the aspect of interaction design in mobile marketing campaigns. Companies who intend to reach the target market effectively should consider factors like interactivity in their marketing materials, what types of technology (Wifi, Bluetooth, RFID, or global positioning system) can the companies adapt in mobile marketing and how best can the mobile medium fit in a given situation as well as to their potential consumers. The technology perspective of mobile marketing should be considered thoroughly.

Third, situational factors like time and location are important issues that any given companies who decide to use the mobile channel for marketing need to consider. At present, there are no concrete solutions and applications for companies to fully adapt when they design their marketing materials.

Fourth, marketing and technology are both at the forefronts of innovation and fashion. Technology introduction-based marketing methods may also drive technology R&D, when the integrated study approach suggests technological improvements.

Lastly, companies may need to consider social factors like culture, values and norms prior to the launch of a mobile marketing campaign. A single set of mobile marketing materials cannot be carried out in different places, as the social factors are often different. Thus, companies operating across many countries may just need to create a general set of guidelines for mobile marketing with the ability to be tailored to specific context situations. The adoption of the Locales Framework is a suitable tool to be considered for such a multi-level guidelines and customisation approach.

CONCLUSION

Using mobile phones as a medium for marketing is a new phenomenon. Companies need to understand the impact of this medium thoroughly before proceeding. Current research into mobile marketing begins with the traditional marketing perspectives and replaces existing media with mobile technology. Such an approach is based on a historical perspective that doesn't arise from the capabilities of human-computer interaction. The introduction of a theoretical-based research framework such as the Locales Framework is suitable in the investigation of this new medium for mobile marketing. A holistic perceptive of technology introduction in a business-to-client interaction to understand mobile marketing is described, with guidelines for supporting the development of mobile marketing strategies. Such a hybrid marketing/interaction design approach generates new possibilities for both technology development and client engagement that either approach individually would not.

ACKNOWLEDGMENTS

This work is supported by ACID (the Australasian CRC for Interaction Design) established and supported under the Cooperative Research Centres Programme through the Australian Government's Department of Education, Science and Training.

REFERENCES

Balasubramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of Academy of Marketing Science, 30*(4), 348-361.

Barnes, S. J. (2002). Wireless digital advertising: Nature and implications. *International Journal of Advertising*, *21*, 399-420.

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing*, *16*(1), 14-24.

Bauer, H. H., Barnes, S. J., Reichardt, T., & Neumann, M. M. (2005). Driving consumer acceptance of mobile marketing: A theoretical framework and empirical study. *Journal of Electronic Commerce Research, 6*(3), 181-192.

Belch, G. E., & Belch, M. A. (2004). *Advertising and promotion: An integrated marketing communications perspective* (6th ed.). Boston: McGraw-Hill/Irwin.

Cheong, J. H., & Park, M-C. (2005). Mobile Internet acceptance in Korea. *Internet Research, 15*(2), 125-140.

Coyle, J. R., & Thorson, E. (2001). The effects of progressive levels of interactivity and vividness in Web marketing sites. *Journal of Advertising*, *30*(3), 65-77.

Fortin, D. R. (2000, June). Clipping coupons in cyberspace: A proposed model of behavior for deal-prone consumers. *Psychology & Marketing*, *17*, 515-534.

Fitzpatrick, G. (1998). *The locales framework: Understanding and designing for cooperative work.* PhD thesis. University of Queensland, Australia.

Fitzpatrick, G. (2003). *The locales framework: Understanding and designing for wicked problems*. Kluwer Academic Publishers.

Fitzpatrick, G., Kaplan, S. K., & Mansfield, T. (1998). *Applying the locales framework to understanding and designing.* Paper presented at OZCHI 1998, Australasian Computer Human Interaction Conference.

Gogus, C. (2004) Understanding young adults' participation in mobile sales promotions. *Paper presented at the 13th EDAMBA Summer School*. Soreze, France.

Haghirian, P., Madlberger, M., & Tanuskova, A. (2005). *Increasing advertising value of mobile marketing—An empirical study of antecedents.* Paper presented at 38th Hawaii International Conference on System Sciences HICSS-38. Hawaii, USA. IEEE Computer Society Press.

Harris, P., Rettie, R., & Cheung, E., (2005). Adoption and usage of m-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom. *Journal of Electronic Commerce Research*, *6*(3), 210-224.

Ito, M., & Okabe, D. (2005). Intimate connections: Contextualizing Japanese youth and mobile messaging. In R. Harper, L. Palen, & A. Taylor (Eds.), *The inside text: Social perspectives on SMS in the mobile age.* London: Kluwer.

Juniper Research. (2004). *M-commerce market to grow to $40bn by 2009*. Juniper Research. Retrieved November 20, 2006, from http://www.finextra.com/fullstory.asp?id=12605

Katz, J. E., & Sugiyama, S. (2005). Mobile phones as fashion statements: The co-creation of mobile communication's public meaning. In R. Ling & P. Pedersen (Eds.), *Mobile communications: Re-negotiation of the social sphere* (pp. 63-81). Surrey, UK: Springer.

Karjaluoto, H. (2005). *An investigation of third generation (3G) mobile technologies and services.* Paper presented at BAI2005 International Conference on Business and Information. Hong Kong.

Kavassalis, P., Spyropoulou, N., Drossos, D., Mitrokostas, E., Gikas, G., & Hatzistamatiou, A. (2003). Mobile permission marketing: Framing the market inquiry. *International Journal of Electronic Commerce, 8*(1), 55-79.

Leppaniemi, M., & Karjaluoto, H. (2005). Factors influencing consumers' willingness to accept mobile advertising: A conceptual model. *International Journal of Mobile Communications, 3*(3), 197-213.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising*, *31*(4), 53-64.

Malloy, A. D., Varshney, U., & Snow, A. P. (2002). Supporting mobile commerce applications using dependable wireless networks. *Mobile Networks and Applications*, *7*, 225-234.

Mayor, T. (2005). *The potential of mobile marketing is huge, but is there more to it than just fun and games?* AlertAds.com. Retrieved November 20, 2006, from http://alertads.com/mobile-marketing-is-huge.html

Okazaki, S. (2004). How do Japanese consumers perceive wireless ad? A multivariate analysis. *International Journal of Advertising*, *23*, 429-454.

Okazaki, S. (2005a). Mobile advertising adoption by multinationals - senior executives' initial responses. *Internet Research*, *15*(2), 160-180.

Okazaki, S. (2005b). New perspectives on m-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-164.

Raghubir, P. J., Inman, J., & Grande, H. (2004). The three faces of price promotions: Economic, informative and affective. *California Management Review*, *46*(4), 1-19.

Rossiter, J. R., & Bellman, S. (2005). *Marketing communications: Theory and applications*. Pearson: Prentice Hall.

Rossiter, J. R., & Percy, L. (1998). *Advertising communications and promotion management* (2nd ed.). The McGraw-Hill Companies, Inc.

Shimp, T. A. (2003). *Advertising, promotion and supplemental aspect to integrated marketing communications* (6th ed.). Thomson: South Western.

Suri, R., Swaminathan, S., & Monroe, K. B. (2004). Price communications in online and print coupons: An empirical investigation. *Journal of Interactive Marketing*, *18*(4).

Tsang, M. M., Ho, S. H., & Liang, T. P. (2004). Consumer attitudes toward mobile advertising: An empirical study. *International Journal of Electronic Commerce*, *8*(3), 65-79.

Wireless Intelligence. (2005). Worldwide cellular connections exceeds 2 billion. *GSM Association Press Release 2005*. Retrieved November 20, 2006, from http://www.gsmworld.com/news/press_2005/press05_21.shtml

Wehmeyer, K., & Müller-Lankenau, C. (2005). *Mobile couponing: Measuring consumers' acceptance and preferences with a limit conjoint approach.* Paper presented at the18th Bled eConference, eIntegration in Action. Slovenia.

KEY TERMS

Computer-Supported Collaborative Work (CSCW): Combines the understanding of the way people work in groups with the enabling technologies of computer networking and associated hardware, software, services and techniques (Wilson, 1991). CSCW also addresses how collaborative activities and coordination can be supported by means of computer systems (Carstensen & Schmidt, 2002). Moreover, CSCW involves incommensurate perspectives as well as incongruent strategies and discordant motives (Schmidt & Bannon, 1992) to gain a better understanding of collaborative efforts within organizations so that this understanding can be used to effectively design collaborative technology that can be best deployed within the organizations.

Human-Computer Interaction (HCI): The study of how people interact with computers and to what extent computers are or are not developed for successful interaction with human beings (ACM SIGCHI, 1996). In fact, HCI is a very broad discipline that encompasses different fields with different perspectives regarding computer development. For instance, HCI in psychology is concerned with the cognitive processes of humans and the behaviour of users, while HCI in computer science is concerned with the application design and engineering of the human interfaces. In sociology and anthropology, HCI is concerned with the interactions between technology, work and organization and the way that human systems and technical systems mutually adapt to each other.

Interaction Design (ID): The study of designing interactive products to support people in their everyday and working lives (Sharp, Rogers, & Preece, 2002). One of the objectives in ID is to produce usable products that are easy to learn, effective to use and also provide an enjoyable experience. Generally users are engaged in the design process.

Locales Framework, The: The Locales Framework is a theoretical-based research framework for interaction design, with a key focus on CSCW. It approaches the study of a certain context or domain with the aim to discover findings that are open in many ways. The general set of guidelines derived from a context or domain is known as the five interdependent aspects. They are: locale foundations, civic structure, individual views, interaction trajectory, and mutuality.

Mobile Advertising (M-Advertising): Advertising is a mass-mediated communication tool. Its aim is to communicate with the intended audience to buy into the desired message. In the context of mobile phones, mobile advertising aims to present the desired information to the consumers, hoping that consumers will react. Currently, most "advertising" contents found on mobile phones are considered as sales promotional materials, which aim to persuade consumers to buy the products.

Mobile Marketing (M-Marketing): Marketing a company's advertised and promotional materials via mobile phones through short message service (SMS) or multimedia messaging service (MMS) is known as mobile marketing. The mobile phone is to the adapted a marketing channel to reach consumers.

Technology Intervention: Technology intervention is the intentional introduction of a technology, or method, into a context to alter that environment. Technology intervention may be targeted at improving information flow, communication, or other types of awareness. In mobile marketing, mobile phones can be seen as a technology intervention in the company-client relationship. Interaction design specifically focuses how to use technology intervention to improve interactions.

3G Commercial Deployment

Mugen Peng
Beijing University of Posts & Telecommunications, China

Shuping Chen
Beijing University of Posts & Telecommunications, China

Wenbo Wang
Beijing University of Posts & Telecommunications, China

INTRODUCTION

Currently, five terrestrial radio interfaces, which can be categorized as frequency division duplex (FDD) and time division duplex (TDD) modes, have been approved as the IMT-2000 radio interfaces. CDMA in TDD (TDD-CDMA) mode will be based on the harmonization between UTRA (UMTS terrestrial radio access) TDD and TD-SCDMA (time division-synchronous CDMA). Compared to UTRA TDD, which is 3.84Mcps in 5MHz bandwidth, TD-SCDMA, is also called low chip rate (LCR) TDD or Narrowband (NB) TDD for its 1.28Mcps in 1.6MHz bandwidth. TDD uses a combined time division and code division multiple access scheme. Hence the signals of different users are separated in both time and code domains (Chen, Fan, & Lu, 2002).

Jointly developed by the China Academy of Telecommunications Technology (CATT) and Siemens, TD-SCDMA is one of the five IMT-2000 standards accepted by the ITU. The main benefits of TD-SCDMA are that it can be implemented less expensively than comparable 3G systems since it is much more spectrum efficient and is compatible with the current deployment of GSM network elements in China, allowing 3G asymmetric services without installation of completely new infrastructure.

Compared with WCDMA and CDMA, TD-SCDMA (Peng et al., 2005) adopts TDD duplex mode, uses the same frequency band for the uplink and downlink, and makes full use of the asymmetrical frequency resource. Meanwhile, the TDD mode has the adjustable switch point between uplink and downlink, which can adapt to the asymmetrical service in uplink and downlink and makes full use of the spectrum. Furthermore, the symmetrical channel feature of TDD systems makes it very flexible and convenient for TD-SCDMA to adopt the advanced technologies such as joint transmission, smart antenna, and so on, which can improve the system capacity and spectrum efficiency. Because the uplink and downlink of TD-SCDMA use the same carrier frequency, the channel propagation features and channel impulse response in downlink and uplink have strong correlation when the interval between uplink reception and downlink transmit is less than the channel coherent time. So the channel information estimated in uplink can be directly used for downlink transmission, and this provides a good condition for implementation of smart antenna. Therefore, compared with the FDD system, channel estimation, power control, and smart antenna in the TD-SCDMA system become more simple and feasible.

As the CDMA system is a self-interfering system, interference among users is the key factor that limits the system capacity (Klein, Irwin, & Roberto, 1991). Using joint detection (JD), inter-symbol interference (ISI), and multiple access interference (MAI) can be effectively eliminated, and as a result, the system capacity is improved (Peng et al., 2004). With the spatial location information, smart antenna can focus the transmit signal power in the direction of the target user, through which users will receive more useful signal power, and interference is highly compressed. This is another way to increase system capacity. However, what is the difference of the radio network planning between TD-SCDMA and WCDMA/cdma2000? Based on the TDD mode, TD-SCDMA covers all application scenarios: voice and data services, packet and circuit-switched transmissions for symmetric and asymmetric traffic, pico, and micro and macro coverage for pedestrian and high mobility users. However, the adopted key techniques, such as time division duplex, smart antenna, multi-user joint detection (MUD), dynamic channel distribution, uplink synchronization, and baton handover, make the network design in TD-SCDMA have a significant different from WCDMA and cdma2000 (Peng et al., 2005). Unfortunately, to the best of our knowledge, there are still no papers investigating and proposing network design solutions for TD-SCDMA in which the impacts of key techniques have been considered. In this article, the key techniques impacting on the radio network design of TD-SCDMA are introduced. Meanwhile, some special radio network design issues for TD-SCDMA are presented.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. TD-SCDMA frame structure

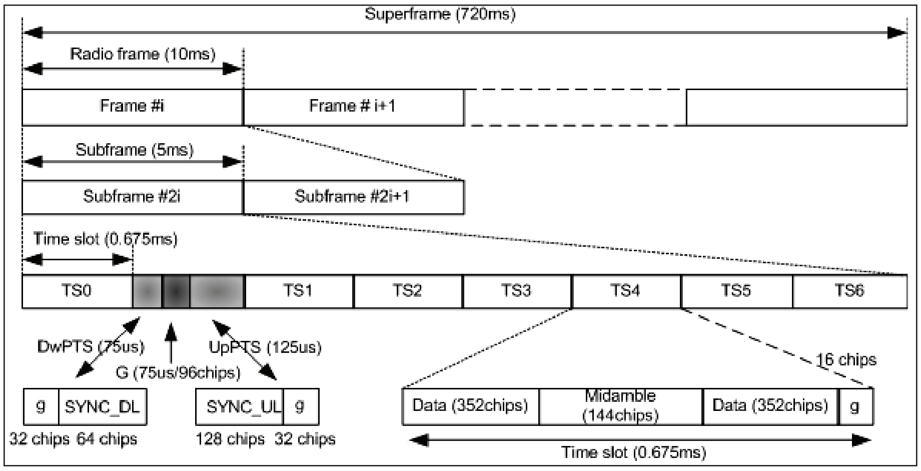
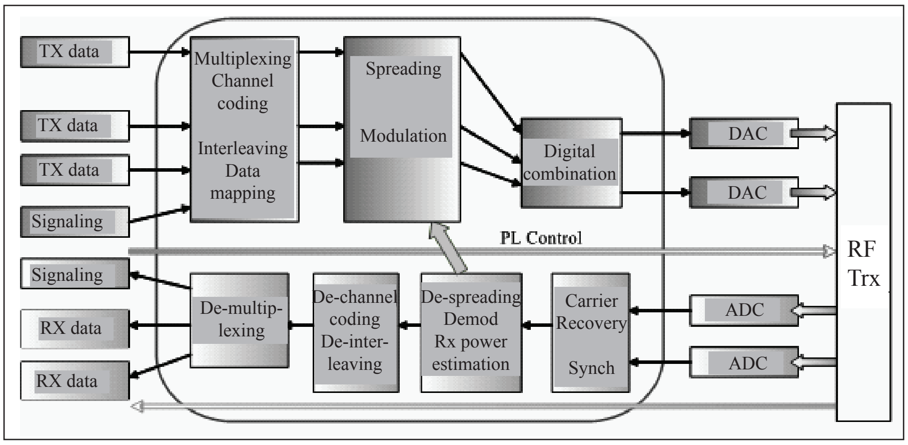

Superframe (720ms)
Radio frame (10ms)
Frame #i
Frame # i+1
Subframe (5ms)
Subframe #2i
Subframe #2i+1
Time slot (0.675ms)
TS0
TS1
TS2
TS3
TS4
TS5
TS6
DwPTS (75us)
G (75us/96chips)
UpPTS (125us)
16 chips
g
SYNC_DL
32 chips 64 chips
SYNC_UL
g
128 chips 32 chips
Data (352chips)
Midamble (144chips)
Data (352chips)
g
Time slot (0.675ms)

Figure 2. System block of TD-SCDMA base band data processing

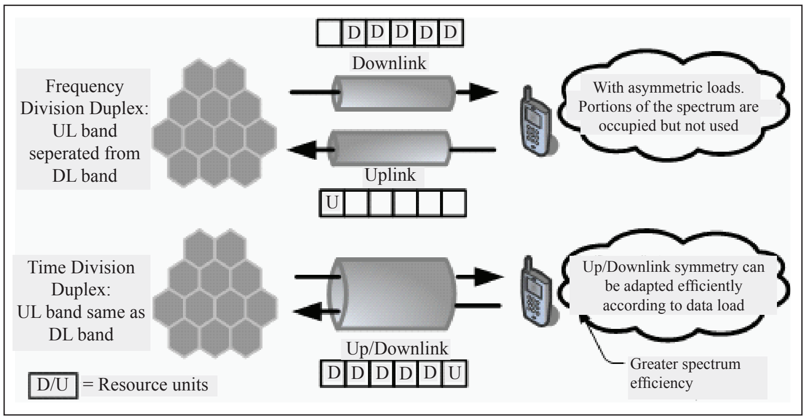

TX data
TX data
TX data
Signaling
Signaling
RX data
RX data
Multiplexing Channel coding
Interleaving Data mapping
Spreading
Modulation
Digital combination
DAC
DAC
PL Control
RF Trx
De-multip-lexing
De-channel coding De-inter-leaving
De-spreading Demod Rx power estimation
Carrier Recovery
Synch
ADC
ADC

SYSTEM BLOCK

TD-SCDMA makes use of both TDMA and CDMA techniques such that channelization in TD-SCDMA is implemented using both time slots and signature codes to differentiate mobile terminals in a cell. The frame structure of TD-SCDMA is shown in Figure 1, where the hierarchy of four different layers—super-frame, radio frame, sub-frame, and time slot—are depicted.

A sub-frame (5 ms) consists of 7 normal time slots and 3 special time slots, where TS0 is reserved for downlink and TS1 for uplink only, whereas the rest (TS2-TS6) should form two groups, the first (whose size can vary from 1 to 4 slots) for uplink and the second (whose size can vary from 4 to 1 slots) for downlink. The slot number ratio of the two groups can take 1/4, 2/3, 3/2, and 4/1 to suit particular traffic requirements. The agility in support of asymmetric traffic is a very attractive feature of TD-SCDMA and of particular importance for Internet services with rich multimedia content in 3G applications. The other three special time slots are downlink pilot (DwPTS), guard period (G), and uplink pilot (UpPTS), respectively. DwPTS and UpPTS are used as a synchronization channel (SCH) for downlink and uplink, respectively, which should be encoded by different pseudo-noise (PN) codes to distinguish different base stations and mobiles. Meanwhile, the subscribers use the midamble part of every burst to estimate the channel impulse response. The subscribers in the same cell are assigned the same basic midamble with different time shift. Using the Steiner estimate principle, the base station can estimate the channel impulse response of all the subscribers simultaneously. The base band data processing is described in Figure 2.

Figure 3. Explaining of TDD and FDD


D D D D D
Downlink
Frequency Division Duplex: UL band seperated from DL band
With asymmetric loads. Portions of the spectrum are occupied but not used
Uplink
U
Time Division Duplex: UL band same as DL band
Up/Downlink symmetry can be adapted efficiently according to data load
Up/Downlink
D D D D D U
Greater spectrum efficiency
D/U = Resource units

KEY TECHNIQUES IN TD-SCDMA

The wireless access technique of TD-SCDMA is based on the TDMA (time division multiple access)/TDD, while WCDMA and cdma2000 are both based on the FDD. Some advanced techniques, such as smart antenna, MUD, and dynamic channel allocation (DCA), can be utilized more conveniently in TD-SCDMA than WCDMA and cdma2000, which can be categorized as the advantage of TD-SCDMA. Meanwhile, the uplink synchronization and handover have a stricter requirement, which can be regarded as the disadvantage in TD-SCDMA.

CDMA Plus TDMA/TDD

The TDD mode allows uplink and downlink on the same frequency band and does not require the pair frequency bands. In TDD, uplink and downlink are transmitted in the same frequency channel but at different times, and the difference between TDD and FDD is described in Figure 3. It is possible to change the duplex switching point and move capacity from uplink to downlink or vice versa, thus utilizing spectrum optimally. It allows for symmetric and asymmetric data services. TDMA is a digital technique that divides each frequency channel into multiple time-slots and thus allows transmission channels to be used by several subscribers at the same time. CDMA increases the traffic density in each cell by enabling simultaneous multiple-user access on the same radio channel. In order to decrease the interference sourcing from the MAI, smart antenna and MUD are utilized. Since TD-SCDMA is based on TDMA/TDD, the air interface for both uplink and downlink is interoperable, which make smart antenna efficiency.

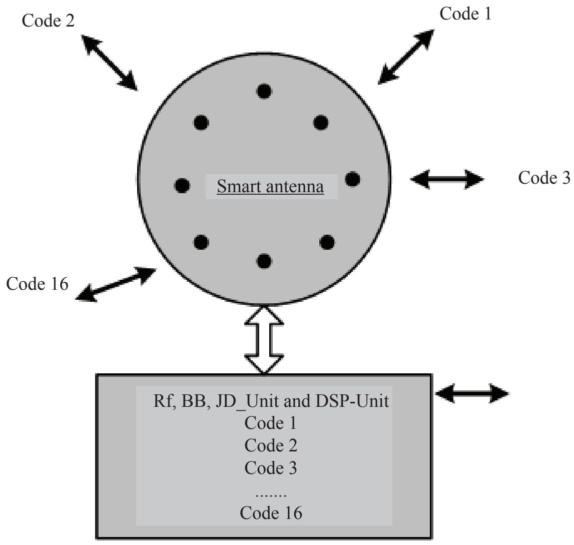
Smart Antenna

A smart antenna system is composed of N antenna elements, N related feed cables, and N coherent radio frequency (RF) transceivers. By using the A/D converters or D/A converters in the analog base-band (ABB), the receiver and transmitter analog signals are interfaced to the digital base-band (DBB) part over the high-speed data bus. The beam-forming which points to a particular user equipment (UE) can be obtained through smart antenna. Due to the inherent robustness of CDMA and the space diversity realized by smart antenna, the interference for multi-path propagation is greatly overcome, and the inter-symbol interface can be greatly reduced. For downlink, beam-forming can also reduce the interference to the other co-channel UEs. These performances can lead to the higher capacity of the TD-SCDMA system.

Smart antennas employed by TD-SCDMA technology are not conventional diversity beam-switching antennas but advanced beam-forming (and beam-steering) bi-directional adaptive antenna arrays. The maximal individual directivity between base stations and mobile terminals is achieved by a concentric array of eight antenna elements with programmable electronic phase and amplitude relations. The terminals tracking is performed by the fast angle of arrival (AOA) measurements in intervals of 5 ms 200 times per second.

The basic idea of the smart antenna is to track the user's mobility, make the interference among different users compressed by spatial filtering, and enhance the desired signal received or focus the energy of the signal on the direction of the desired user location. Thus the system coverage, power efficiency, and system capacity can be improved. The most common antenna array geometry structures are uniform linear array (ULA) and uniform circle array (UCA), which is described in Figure 4.

Figure 4. Basic principle of uniform circle array



The beam pattern of the ULA is symmetrical along the boresight of the antenna array, that is to say ULA cannot distinguish the users that locate at the symmetry location along the boresight of the antenna array. The more the user location deviates the boresight of the antenna array, the wider the beam and the lower the resolution of the antenna array will be.

For UCA, as the distance between the elements increases, the aperture of the antenna array becomes larger, and the resolution of AOA increases; meanwhile the amplitude of the side lobe becomes larger. Compared to the ULA, no phase blur happens. So except the sector scenario, UCA will be adopted.

Three types of beam-forming antennas are introduced in TD-SCDMA: phased array, switch-beam, and Eigen-beamforming. For the phased array, the weighted combining only do some phase rotation on the received signals. Switch-beam is a simplification of the phased array, in which the whole cell is divided into several sectors and a set of pre-defined beam-forming weights are used to cover these sectors one by one. The rationale of the Eigen-beamforming is that the maximum Eigen value and Eigen vector of the spatial correlation matrix of the user is found by Eigen value decomposition as the power gain and beamforming weight. The performance of the switch-beam is little worse than that of phased array, and the Eigen-beamforming is expected to be the best.

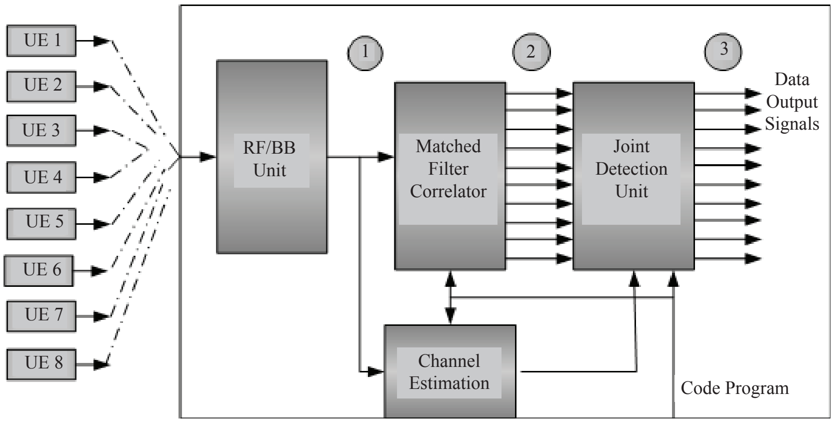
In the current TD-SCDMA system, the phased array smart antenna is utilized to suppress the interference from both the inter-cell and intra-cell. However, if the air condition is too bad and the AOA estimation is incorrect, the smart antenna cannot work efficiently. In this way, the multi-user joint detection technique is adopted to suppress only the intra-cell interference.

Multi-User Joint Detection

Multi-user joint detection (MUD) allows the receiver to estimate the radio channel and works for all signals simultaneously. Through the parallel processing of individual traffic streams, MUD eliminates the MAI and minimizes intra-cell interference, thus increasing the transmission capacity.

Figure 5 describes the basic principle of MUD. The first step is that all signals from the carious UEs are received in the node B receiver through the radio frequency (RF)/baseband (BB). The second step is to detect the signals barely emerging from the MAI with a low signal to noise ratio (SNR). In the last step, using a specific algorithm, a DSP thus extracts all CDMA channels in parallel and removes the interference caused by the undesired CDMA channels. The result is a clear signal (high signal to noise ratio) for each CDMA code.

Figure 5. Basic principle of MUD



Note that the efficiency of MUD in TD-SCDMA technology is based on the TDMA/TDD operation and on the limited number of codes employed. The total number of users per radio carrier is distributed over the different time slots of the basic TDMA frame, so that a maximal number of 16 codes per time slot per radio carrier can be easily processed in parallel and detected. However, due to the huge number of spread codes used by WCDMA and cdma2000, the implementation of an optimal multi-user receiver in these systems is difficult, since the implementation complexity is an exponential function of the numbers of codes. In order to combat MAI, both WCDMA and cdma2000 systems employ the suboptimal detection schemes, such as the rake receiver, which do not extract all CDMA codes in parallel.

Dynamic Channel Allocation

A further minimization of inter-cell interference is achieved by dynamic channel allocation (DCA). There are different radio resource dimensions for TD-SCDMA: TDMA, FDMA, CDMA, and SDMA (space division multiple access). Making an optimal use of these degrees of freedom, DCA provides an adaptive allocation of the radio resources according to the interference scenario, minimizing the inter-cell interference (Peng et al., 2003).

In TD-SCDMA, the DCA is divided into two parts: Slow DCA allocates resources to cells, while Fast DCA allocates resources to the bearer services, balances the load in the different slot/frequency, and congregates the radio resource for supporting the high bit rate services. Both UEs and node Bs perform the periodic monitoring and reporting to support DCA. Fast DCA is always terminated at the node B, but slow DCA can be terminated at any network entity above the node-Bs that forms the seamless coverage area. The slow DCA algorithm allocates the radio resource units in a cell-related preference list for Fast DCA to acquire them for different bearers. In the first phase the cell-related preference list is a fixed table that is given as a parameter for each base station. Fast DCA resides in each node B and is responsible for utilizing the slots assigned to the node B in the most efficient manner possible. This involves: (1) assigning each UE to the slot(s) best suited in each particular case at the start of a connection, and (2) reshuffling/reallocation of UEs when traffic and/or environmental conditions have changed or if the requirements of an existing connection have been altered by the UE.

According to the asymmetry of transmission bits, the switched point between uplink and downlink is different in the adjacent cells in which the cross slot interference sourcing from the node B-node B is huge. In order to suppress the cross slot interference, the Fast DCA combines the smart antenna technique to allocate the radio resource in the cross slots to the UEs close to the serving node B. In this way, the total transmission power in the cross slots decreases and the interference is suppressed.

Uplink Synchronization

Like all TDMA systems (e.g., GSM), TD-SCDMA needs an accurate synchronization between mobile terminal and base station—that is, the UEs' spread signals arriving at the node B at the same time can effectively simplify the demodulator in node B, decrease MAI, and increase the system capacity. This synchronization becomes more complex through the mobility of the subscribers, because they can stay at varying distances from the node B, and their signals present varying propagation times.

An uplink synchronization procedure includes two stages: synchronization establishment and synchronization maintenance. The synchronization establishment is often associated with UE's access procedure, and the synchronization maintenance is often associated with the dedicated communication procedure between UE and network. Due to the multipath and shadow fading, however, the establishment and maintenance of the uplink synchronization between different UEs in a cell are difficult. The simple, effective, and low-cost establishing and maintaining uplink synchronization solution will bring benefits to TD-SCDMA. The UL synchronization is equivalent to a very high precise timing advance according to the physical layer specifications. Therefore, an extended time advanced (TA) option by means of a sub-chip granular operation is utilized in TD-SCDMA. The granularity of TA, in the case of UL synchronization, is ±1/8 chips.

Baton Handover

The conventional hard handover has some shortages, for example the dropping ratio of the hard handover is high, and the efficiency of the radio resource usage of soft handover is low. With smart antenna and uplink synchronization, the handover strategy will change and the performance will be improved in TD-SCDMA. The UE position information and AOA are provided to predict the handover request, prepare for handover and pre-synchronize for handover, shorten handover time, decrease handover blocking, simplify handover procedure, and improve handover confidence. This is named baton handover in TD-SCDMA (Peng et al., 2003).

The basic principles of baton handover are: (1) the system knows the position of all UE, (2) the system knows and determines the target cell for handover, (3) the system informs mobile the information about the node B in neighboring cells, (4) mobile measurement helps the system to make the final decision, and (5) after the cell search procedure, the mobile UT has already established.

Since UE can synchronize to the node-B in target cell, the baton handover costs a shorter handover time period for both inside the TD-SCDMA system and between TD-SCDMA and the different systems. In the TD-SCDMA system, the parameters that UE measure are not only the received signal power level, but also their transmission time offset and so on.

TD-SCDMA RADIO NETWORK DESIGN

Since the TD-SCDMA system adopts many advanced techniques, there are some unique issues when doing TD-SCDMA radio network designs (Peng et al., 2005). First, the scramble codes in TD-SCDMA are only 16 chip lengths, and the orthogonality between different scramble codes is easy to lose which will cause difficulty in pilot searching and reception. So, the scramble code planning is especially important in a TD-SCDMA system. Second, as TD-SCDMA is a narrowband system, its bandwidth is one-third of that of WCDMA, and the capacity of single carrier TD-SCDMA is low compared to WCDMA. That means if TD-SCDMA occupies the same bandwidth as WCDMA, the multi-carrier system can be adopted to form a higher capacity (Peng, Chen, & Wang, 2006). Besides, multi-carrier with careful configuration of carrier can make scramble codes planning much easier. Multi-carrier TD-SCDMA layout is proposed to form a network and provide many kinds of mobile services. Third, coverage of a certain system is of great importance. It can be analyzed in a theoretical way by link budget. Link budget of TD-SCDMA differs from that of CDMA2000 and WCDMA, due to its own properties and the advanced techniques that TD-SCDMA employs. Finally, coexistence of TD-SCDMA with other 3G systems is another important issue both for TD-SCDMA and other related system performances (Peng et al., 2004).

CONCLUSION

In TD-SCDMA only 16 codes for each timeslot for each carrier are used. The intra-cell interference is eliminated by MUD, and inter-cell interference is minimized by the joint use of smart antennas and DCA. Meanwhile, since most interference is suppressed under the condition of smart antenna working efficiently, the capacity is radio resource limited, and the cell breathing effect is not an issue anymore. However, some new problems occur due to the key techniques and features of TD-SCDMA, such as the N frequency planning, scrambling code planning, multi-operators coexistence planning, and advanced radio resource management.

In order to successfully deploy the commercial TD-SCDMA network, the key techniques impacting on the network performance and the novel network strategies should be proposed. The special network planning method and project steps should be configured for TD-SCDMA. This article discussed all these confusing problems in TD-SCDMA network planning and presented the principle solutions. Some special issues should be investigated, such as the handover strategies between TD-SCDMA and other mobile communication systems, indoor network design, and repeater design specified for TD-SCDMA.

REFERENCES

Chen, H. H., Fan, C. X., & Lu, W. W. (2002). China's perspectives on 3G mobile communications and beyond: TD-SCDMA technology. *Wireless Communications, 9*(2), 48-59.

Klein, S. G., Irwin, M. J., & Roberto, P. (1991). On the capacity of a cellular CDMA system. *Vehicular Technology, 40*(2), 303-312.

Peng, M. G., Bao, W., Hu, W., & Wang, W. B. (2004). Investigation of uplink admission control schemes for TDD-CDMA systems. *Proceedings of the International Conference on Communications, Circuits and Systems* (pp. 443-446). Chendu, China.

Peng, M. G., Chen, S. P., & Wang, W. B. (2006). TD-SCDMA evolution and multi-carrier techniques. *Telecommunication Science, 22*(5).

Peng, M. G., Hu, W., & Wang, W. B. (2004). Investigation of uplink capacity based on the background noise floor in TDD-CDMA systems. *Proceedings of the International Conference on Signal Processing* (Vol. 3, pp. 1918-1921). Beijing, China.

Peng, M. G., Huang, B., & Wang, W. B. (2004a). TDD-CDMA capacity loss due to adjacent channel interference in the macro environment employing smart antenna techniques. *Proceedings of the 2004 Asia-Pacific Radio Science Conference* (pp. 146-149). Qingdao, China.

Peng, M. G., Huang, B., & Wang, W. B. (2004b). Investigation of TDD and FDD CDMA coexistence in the macro environment employing smart antenna techniques. *Proceedings of the 5th International Symposium on Multi-Dimensional Mobile Communications, 2004 Joint Conference of the 10th Asia-Pacific Conference* (Vol. 1, pp. 43-47). Beijing, China.

Peng, M. G., & Wang, W. B. (2003). Novel approaches for downlink performance analysis in CDMA networks. *Proceedings of PIMRC 2003* (Vol. 3, pp. 2533-2537). Beijing, China.

Peng, M. G., & Wang, W. B. (2004a). Advanced HARQ and scheduler schemes in TDD-CDMA HSDPA systems. *Proceedings of the 5th International Symposium on Multi-*

Dimensional Mobile Communications, 2004 Joint Conference of the 10th Asia-Pacific Conference (Vol. 1, pp. 67-70). Beijing, China.

Peng, M. G., & Wang, W. B. (2004b). An analysis of resource allocation and management in TDD-CDMA systems employing smart antennas. *Emerging Technologies: Frontiers of Mobile and Wireless Communication, 2,* 753-756.

Peng, M. G., & Wang, W. B. (2004c). TDD-CDMA uplink capacity investigation in the background noise floor. *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 233-236).

Peng, M. G., & Wang, W. B. (2005a). A framework for investigating radio resource management algorithms in TD-SCDMA systems. *IEEE Communication Magazine, 43*(6), 12-18.

Peng, M. G., & Wang, W. B. (2005b). Comparison of capacity between adaptive tracking and switched beam smart antenna techniques in TDD-CDMA systems. *Proceedings of Mape2005,* Beijing, China.

Peng, M. G., & Wang, W. B. (2005c). Investigation of handover strategies in TDD-CDMA cellular networks. *Proceedings of the ACM SIGCOMM Asia Workshop 2005,* Beijing, China.

Peng, M. G., & Wang, W. B. (2005d). Investigation of the distributed antenna scheme for multi-cell environment in TDD-CDMA systems. *Proceedings of ITC2005,* Beijing, China.

Peng, M. G., & Wang, W. B. (2005e). Investigation of uplink performances based on the switched beam antenna scheme in TDD-CDMA systems. *Proceedings of ITC2005,* Beijing, China.

Peng, M. G., & Wang, W. B. (2005f). *TD-SCDMA mobile communication system.* Beijing: China Machine Press.

Peng, M. G., & Wang, W. B. (2005g). TD-SCDMA network planning. *Telecommunication Technology, 5,* 6-9.

Peng, M. G., Wu, Y. C., & Wang, W. B. (2004, May 17-19). Joint and advanced proportionally fair scheduling and rate adaptation for multi-services in TDD-CDMA systems. *Proceedings of VTC 2004* (Vol. 3, pp. 1630-1634). Milan, Italy.

Peng, M. G., Zhang, J. W., Hu, C. J., & Wang, W. B. (2003). Handover performance analysis in TDD-CDMA cellular network. *Proceedings of WCNC 2003* (Vol. 2, pp. 806-811).

Peng, M. G., Zhang, J. W., Liu, Y., & Wang, W. B. (2003). On the capacity of a cellular TDD-CDMA system employing frequency channel assignment schemes. *Proceedings of the International Conference on Communication Technology* (Vol. 2, pp. 803-807). Beijing China.

Peng, M. G., Zhang, J. W., Zhu, X. M., & Wang, W. B. (2003). A novel dynamic channel allocation scheme to support asymmetrical services in TDD-CDMA systems. *Proceedings of the International Conference on Communication Technology* (Vol. 2, pp. 794-798). Beijing, China.

KEY TERMS

Code Division Multiple Access (CDMA): A kind of multi-access method. Users are distinguished in code domain. Other well-known multiple access methods are FDMA (frequency division multiple access), in which frequencies are used to distinguish different users; and TDMA (time division multiple access), in which time is used to distinguish different users.

DCA: Another key technique that a TD-SCDMA system holds, by which radio resources are allocated to different traffic bears of different cells.

Multiple Access Interference (MAI): The main kind of interference in a CDMA system. It has a bad impact on CDMA system capacity.

Multi-User joint Detection (MUD): A kind of detection method at receiver. Other common detection methods are rake receiver and matching receiver.

Radio Network Design: Design of network framework, layout, and so on. This is a key for a system to be commercially feasible.

Smart Antenna: A key technique of a TD-SCDMA system which can enlarge received signal power and suppress interference, thus improving received SNR (signal to noise ratio) and enlarging system capacity.

System Capacity: The number of users of certain traffic that the system simultaneously holds. Other definitions are the number of Erlang that a system holds simultaneously.

Time Division Duplex (TDD): A kind of duplex mode. Another well-known duplex is FDD (frequency division duplex) mode. Uplink (from user to network) and downlink (from network to user) in TDD mode are divided by time and in FDD by frequency.

Time Division duplex-Synchronous Code Division Multi-Access (TD-SCDMA): A 3G standard proposed by the China Academy of Telecommunications Technology (CATT) and Siemens jointly, and accepted by the ITU in 1999.

Transaction Management in Mobile Databases

Ziyad Tariq Abdul-Mehdi
Multimedia University, Malaysia

Ali Bin Mamat
Universiti Putra Malaysia, Malaysia

Hamidah Ibrahim
Universiti Putra Malaysia, Malaysia

Mustafa M. Dirs
College University Technology Tun Hussein Onn, Malaysia

INTRODUCTION

Recent advances in wireless communications and computer technology have provided users the opportunity to access information and services regardless of their physical location or movement behavior. In the context of database applications, these mobile users should have the ability to both query and update public, private, and corporate databases. The main goal of mobile software research is to provide as much functionality of network computing as possible within the limits of the mobile computer's capabilities. Consequently, transaction processing and efficient update techniques for mobile and disconnected operations have been very popular. In this article, we present the main architecture of mobile transactions and the characteristics with a database perspective. Some of the extensive transaction models and transaction processing for mobile computing are discussed with their underlying assumptions. A brief comparison of the models is also included

TRANSACTION MANAGEMENT IN MOBILE DATABASES

A mobile database system is a special multi-database system on a mobile computing environment. It allows mobile hosts to access and manipulate data stored on several per-existing, autonomous, and heterogeneous local database systems located on different parts of the wired network. Transactions in a mobile database system may access data from several local databases at different sites. Management of these transactions requires different approaches in mobile databases than in a multi-database. This is mainly due to the fact that a mobile host is not suitable to manage a global transaction by itself due to the described nature of the mobile computing environment. Usually this management is done by the mobile host's base station or by coordination of it.

Due to the described nature of the mobile computing environments, transaction management has to be reevaluated for mobile databases. The transactions in mobile computing environments are usually long-living transactions, possibly covering one or more disconnected durations. Supporting disconnected operation (i.e., allowing a mobile host to operate autonomously during disconnection) raises issues in consistency. Providing disconnected operation also requires some pre-caching of data that will be required for the necessary operations to be performed during disconnection. The moving behavior of the transactions in mobile computing environments also requires new mechanisms. As a mobile host moves from a cell to another cell, its transactions might need to migrate from one base station to another. In general, transactions in mobile databases require relaxed ACID properties. There are several works on mobile transactions, each addressing some of the issues in mobile transaction management. We will explain some of them in the following sections.

Kangaroo Transactions

Kangaroo transactions (KTs) are introduced in Dunham and Helal (1997). As the name suggests, this model mostly addresses the moving behavior of the mobile transactions. As the transactions hop from one site to another, the management of the transaction also moves.

In addition to the mobile computing environment we have described, these systems introduce a couple of other terminologies. The term *source system* represents a collection of systems that offer information services to mobile users. These systems could be any type of system that exists in the mobile computing environment. One good example is a distributed database system. The term *data access agent (DAA)* represents an agent that is hosted by each base station. Mobile hosts reach data in source systems by sending their transactions to DAAs. When a handoff occurs, the DAA at

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the new base station receives the transaction information from the DAA in the old base station. A *mobile transaction* is defined as the basic unit of computation in the mobile environment. The management of a mobile transaction might hop through different base stations, which are not known until it completes its execution. DAAs at base stations are responsible for management of the mobile transactions. One part of the DAA responsible for the management of the transactions is called the *mobile transaction manager (MTM)*. The main responsibilities of an MTM are maintaining the status of mobile transactions in execution, logging recovery information, and performing needed check pointing.

In this model it is assumed that a mobile transaction issued by a mobile host to a DAA might include several subtransactions that require access to data at several global database systems (GDBSs) and DBMSs residing at different places of the fixed network. As a result DAAs serve as a mobile transaction manager built on top of GDBSs and DBMSs. DAAs also keep log information about the mobile transaction parts that have executed on them. Remember that a mobile transaction changes its DAA as it moves from one cell to another.

Since mobile transactions are long lived and include possible disconnected durations; the atomicity of a mobile transaction in this model is not always guaranteed. The time between an interruption of a transaction and its resume could be quite long. As a result, it is valuable to commit early on some portions of a mobile transaction, while breaking the atomicity property of the transactions. These early commits enable the release of possible important resources, instead of holding them for a long time.

A mobile transaction in this model, which is called a *kangaroo* transaction, is an extension to global transactions (GTs). Figure 1 shows a global transaction, which consists of subtransactions called local transactions (LTs). Each local transaction is assumed to be issued to a DMBS.

A kangaroo transaction can be composed of both GTs and LTs. The mixture of GTs and LTs are grouped under a transaction type called Joey transactions (JTs) based on the DAA on which they have initiated. Figure 2 shows an example kangaroo transaction.

Each JT represents the unit of execution at one base station. When a mobile host makes a transaction request to the DAA on its associated base station, a KT is formed. In addition to that, a JT is formed for managing subtransactions that originate from the mobile host when the KT is under the control of the first DAA. When a mobile host hops from one cell to another, the control of its KTs changes to the new DAA on the new base station. The new DAA creates a new JT for handling the future subtransactions the mobile host might request to this DAA. The old JT is committed independently from the new one. Note that this breaks atomicity. To enable a KT to be completely undone, previously committed transactions should be compensated.

Kangaroo transactions have two different modes of execution. The first mode is called *compensating mode* where a JT falls and all KT is undone. However, this mode of operation requires compensating transactions for undoing operations of the previous JTs, since they are independently committed building compensating transactions requiring input from the user. As a result this mode is rarely used. Note that this mode tries to preserve atomicity of the KT, which breaks the durability of the subtransactions. The second mode, which is the default mode, is called the *split mode.* In this mode of operation, when a JT fails, no new JTs are created, but the previously committed Ms are also not undone. This mode breaks the atomicity. None of the modes ensures serializability.

Clustered Model

In Pitoura and Bhargava (1995), the sites are grouped in clusters if they are connected by strong network links or they are pre-grouped in clusters. In a cluster, full consistency is always enforced, however over the clusters a bounded inconsistency is permitted. A site can be a member of a cluster or leave one dynamically based oil the network conditions, for example, a mobile host forms a cluster itself when it is disconnected.

Mobile transactions are grouped into two types: strict and weak transactions. Weak transactions access data in the same cluster, and they have two commit points: cluster and global. Global commit can only be made after clusters merge. Strict transaction can only access data that is ac-

Figure 1.

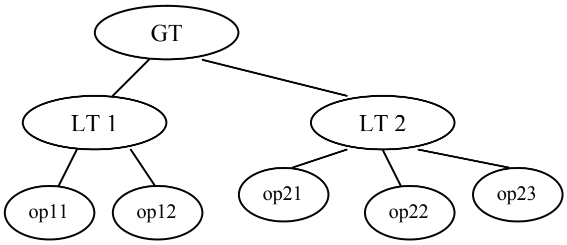
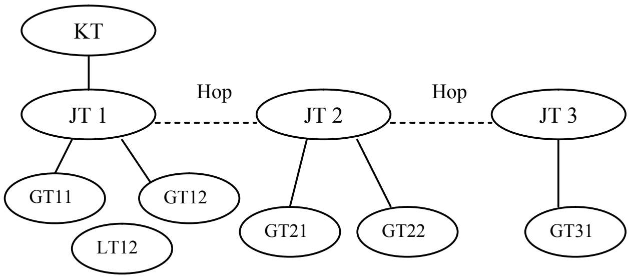


Figure 2.



cessed by a strict transaction or globally committed weak transaction. They are not allowed to access inconsistent data. Applications updating private data-data—mostly updated by the mobile host on itself—can use weak transactions during disconnections. Applications accessing global data can use weak transactions also if they are tolerant of inconsistent data to some level.

A Pre-Serialization Transaction Management Model

This model (Dircke & Gruenwald, 2000) also assumes a fixed and attached mobile network like the kangaroo model. MUs connect to MSSs. MUs access the mobile multi-database system (MMDBS) via an interface called global transaction manager (GTM). GTM is responsible for providing consistent and reliable units of computing. GTM is a group of global transaction coordinators (GTCs) and site transaction managers (STMs). A GTC runs on every MSS. There exists an STM on each local database. GTC handles disconnections and logging, and delivers the transactions and transaction compensators to the STMs. The STM returns the commit or abort results to the GTC once the transaction submitted to it is completed. In case of a global abort, the compensators on STMs are used to roll-back the transactions. GTC tries to forward the result to the MU or saves it if the MU is disconnected. GTC also tries to conclude if an MU is disconnected for a short time or failed for a prolonged period. Transactions of MUs that are failed are aborted as they become obstacles to other transactions.

A transaction is a group of compensatable, open-nested, vital and non-vital subtransactions. All subtransactions can commit independently. After all the vital subtransactions are committed, a transaction is checked for atomicity and isolation (A/I). After this point if that transaction is not aborted, it can submit only non-vital subtransactions. If any vital subtransaction aborts, then the transaction is aborted.

An algorithm called partial global serialization graph (PGSG) algorithm is used to check the A/I constraints. Each STM maintains a local serialization graph for vital transactions. These graphs are forwarded to GTCs upon request. After merging these graphs if there is no cycle then the transaction is toggled (marked checked), else it is aborted and all subtransactions are compensated.

A toggled transaction is guaranteed to commit if its mobile station is alive or only disconnected for a small amount of time. If a catastrophic failure has occurred, then a toggled transaction can be aborted if it obstructs other transactions.

This model allows subtransactions to commit independently, thus allowing them to release resources as they are not needed. Also a suspended state where transactions can be aborted is introduced. In case of a catastrophic failure, an MU is assumed to be in the suspended state. However, determining if an MU has failed should be done carefully, so not to abort transactions of a disconnected host.

Deno

In Cetintemel and Kelender (2002), Deno, a replicated object storage system designed for use in mobile and weakly connected environments, is discussed. Deno is designed to support weak connections and limitations of mobile hosts, such as limited processing power and limited coverage area. Thus, it is a lightweight, peer-to-peer, fully decentralized, and asynchronous system.

An MU does not need to know the other hosts, but it needs to be in contact with at least one other node. Pair-wise anti-entropy sessions are used to spread information—votes in this case; update transactions are always voted to commit after they are committed locally on a node. This scheme works as follows: if two nodes x and y can communicate with each other, x can ask y if it knows any globally committed transactions it does not know, if so it copies this information. If y does not know anything, it just sends commit candidates it voted for to x and casts its own vote. Any voter keeps track of a number of votes for an object k and the number of unknown votes. Since currencies (weights) are distributed such that the total is 1.0, any node can keep track of unknown votes. If on any MU the votes for k are more than the votes for j and unknown, then it can commit k. This method ensures that all updates are committed in the same order globally.

In the case of planned disconnections (e.g., sleep mode), a node can appoint a proxy for itself and transfers its currency to that proxy. After returning to the network, the formerly disconnected node can claim its currency back.

In case of an unplanned disconnection, a proxy for the disconnected node is selected using the voting scheme described above. If the proxy election is globally committed, the currency (weight) of the disconnected node is transferred to the elected proxy. In the case of network partitioning, only proxies for the partition that has the smaller total currency can be selected because the proxy election is done using the voting scheme described above. As in the planned case, any re-connected MU can claim its currency.

Weight assignment is crucial to designing a good working scheme; basically, assigning more weights to strongly connected hosts (e.g., stable and more powerful hosts) is a good strategy. Currency distribution is discussed in detail in Cetintemel and Kelender (2002).

Pro-Motion Model

The pro-motion model (Walborn & Chrysanthis, 1996) is designed to support disconnected transaction processing. The motivation is that disconnected MUs can execute transactions if they have the data and methods required.

The fundamental building block is the compact, which is the basic unit of data replication for caching and hoarding. A compact is not only a piece of data, but it is an object that includes restrictions (allowable operations), obligations (such as an expiration time), and methods. In other words compact is a mini database that is moved to the local hosts upon request. A database server delegates an MU as the controller of a compact for local transaction processing. The MU should agree on all obligations and restrictions that the database server sets by the database.

Pro-motion uses open-nested split transactions. When an MU is connected to the network, it identifies a group of compacts that are updated by locally committed subtransactions. Those transactions are split from the uncommitted ones and sent to the owner of the compact. Those transactions are then committed on the database, making the updates visible to all other transactions.

An MU first caches the compacts it needs, then disconnects and processes the transactions. When reconnected it resynchronizes with the fixed database. Transactions can be of two types—local and traditional. Local transaction results are made visible when they commit to the other transactions on the same MU. A transaction can also be traditional, which means its results are invisible until resynchronization.

To allow the resources to be released in a timely manner, each compact is assigned a deadline. An MU can request an extension if the deadline has passed. If the compact is free, then a new deadline is negotiated; if not, this compact is marked invalid and all transactions accessed are aborted. Also the other compacts written by these transactions are marked unavailable, and transactions read from unavailable compacts are aborted.

The valid compacts are resynchronized with the databases. The operations made on the compact are sent to the database servers. These operations are executed as a single transaction. If this transaction can complete, then a commit message is returned to the MU, else the MU receives an abort message and executes compensating procedures for the committed local transactions' owners.

A 10-level scale of correctness is defined in pro-motion from serial to no-guarantee. Each method in a compact is assigned a level. Each transaction is also assigned a minimal level for READ and WRITE operations, and it can only perform operations at that level and READ operations on a higher level than WRITE operations.

TCOT Protocol

Kumar, Prabhu, Dunham, and Seydim (2002) present a timeout-based commitment protocol, TCOT. Like Deno, it is designed for weakly connected, less powerful MUs. TCOT tries to minimize the communication, since bandwidth is scarce and is non-blocking since MUs can disconnect unpredictably. It assumes that MU has a cache and transaction processing power.

A coordinator (CO) is responsible for a transaction submitted by an MU. COs are either base stations (BSs) or a node on the fixed network. The MU extracts the subtransaction t_i,j; it will run from transaction $T_i = \{t_i,j | l \leq n \leq n\}$ and sends $T_i - t_i,j$ to the CO. MU sends extra information such as how long it will take to process the subtransaction and to ship the updates to the CO. The CO distributes the subtransaction to the fixed nodes on the network and starts its timer according to the values it received from the MU.

Any node executing a subtransaction can request a time extension. But if the CO does not receive a message from the MU or the other participating nodes, then it aborts the transaction and sends an abort message to all of the nodes. Else, if it receives commit messages from all nodes and updates from MU, it does not send any further messages.

Since some nodes commit without a global commit, compensating transactions are necessary to undo the globally aborted but locally committed subtransactions.

Choosing or calculating timeout values is important through this scheme; extension messages are minimized and the transaction restar MUs should be able to choose values for initial timeout values based on the network conditions. Moreover they can request extensions from those incremented with each request. The CO should be able to reject an extension request if the system throughput falls tinder-desired value.

MANET Model

A mobile ad-hoc network (MANET) has restrictions, which make models like TCOT and kangaroo transactions infeasible to use. MANETs lack nodes which are on fixed networks and BSs which can act as CO or DAA. Moreover, since all nodes route traffic, if they fail the network starts to weaken, partitions can occur, and usable bandwidth between the nodes starts to drop.

Gruenwald and Banik (2001) propose a model to tackle these problems. It is assumed that there are two types of nodes, large mobile hosts (LMHs) and small mobile hosts (SMHs). LMHs have more processing power, storage capacity, and power source than SMHs. An SMH sends its transaction to an LMH depending on the transaction type. If the transaction is a firm transaction, it should be finished before its deadline or it has no value. If the transaction is soft, then it has two deadlines. The earlier deadline can be violated, but the value of the transaction starts to decrease to zero towards its second deadline.

A fixed transaction is always processed by the nearest LMH, whereas a soft transaction is processed by the LMH with the highest remaining energy level. LMH distributes the subtransactions to other LMHs upon submission. The subtransactions are also grouped into two sets: *vitals* and *non-vitals.* After all vital subtransactions finish, the transaction is verified against atomicity and isolation using PGSG (Dircke & Gruenwald, 2000), as described earlier. If the transaction does not violate these conditions, it would be marked as ready and toggled to commit, and it starts to wait for the non-vital subtransactions.

A toggled transaction is guaranteed to commit unless it blocks another global transaction while it is in the suspended state. A transaction is in suspended state if its MU disconnects.

The LMH tries to send the result to the originating the SMH, but if it fails while sending the result of a firm transaction, it aborts the transaction. However, if it fails while trying to send a soft transaction, it retries until the second deadline is reached.

While a transaction is executing, the SMH can be in doze mode or sleep mode to conserve energy. If it is waiting for a result of fixed transaction, it should not go into the sleep mode for a long time, otherwise the transaction would be aborted. If it is waiting for a soft transaction, it can sleep until the end of the second deadline, saving energy.

Planned Disconnection Modes

Planned disconnection modes (Datta et al., 1999; Demers et al., 1994) involve informing the distributed system of the intention to disconnect and may include the appointing of a proxy. The purpose of a planned disconnection procedure is to enable the remaining connected sites to continue processing with minimal disruption. There are a number of different ways that a disconnection can affect the database. In this section, we explore the possible kinds of planned disconnection and provide appropriate terminology.

In *basic sign-off* mode (Datta et al., 1999), an MH decides to disconnect and informs the system, consisting of the currently connected sites, of its intention. The database of the disconnected MH becomes read-only while the access capabilities of the remaining connected sites are unaffected.

In *check-out* mode (Demers et al., 1994), the MH wants to disconnect and be able to update a set of data items X. There are three variations to this mode that determine what type of access to non-checked-out items is allowed.

The first variation is *DB partition.* In DB partition mode, the database is partitioned into *X* and *DB−X*. The disconnected site has complete and unlimited access to *X* and nothing else, while the remaining system has complete access to *DB−X* and nothing else. The second variation is *check-out with mobile read.* This mode allows the disconnected site to have read access to all database items in addition to the read/write access to the checked-out items *X*. The remaining connected sites in the system have complete (read/write) access to *DB−X* and no access to *X*.

The third variation is *check-out with system read.* This mode allows the connected sites in the system to have read access to all database items in addition to the read/write access to the non-checked-out items *DB−X*. In check-out mode, when the MH checks out an object, either the MH or the remaining connected sites are prevented from accessing to read some of the objects in the database. Although this is necessary to preserve serializability, many database systems operate on lesser degrees of isolation (Chrysanthis et al., 1994). Therefore, we define a *relaxed check-out* mode in which the remaining sites can read the items that other sites have checked out, while disconnected sites can read items they have not checked out.

COMPARISON OF MODELS

Kangaroo, pre-serialization, TCOT, and pro-motion models are designed for multi-tiered networks. It is assumed that there is a fixed and reliable wired network that can support mobile hosts. Moreover, all local databases are assumed to be located in the fixed network, which makes replication management and deadlock detection algorithms easier to implement.

These models mentioned above are all making use of a similar notion called DAAs. DAAs execute the transaction for mobile units, and since they reside on the fixed network, they are never disconnected.

Pro-motion is different from the other terms in that mobile units can execute transactions by themselves, which is an advantage if they are working on small units of data and they need user interaction. In other systems user interaction is not considered. If a transaction needs user input, it should wait for the originating mobile node to be disconnected. A drawback of pro-motion is that it assumes that wireless bandwidth is not a scarce resource, since it moves data back and forth.

Kangaroo is the only modal that addresses the mobility issue. Although in the other models mobility can be handled similarly to kangaroo, it is not explicitly addressed.

In the models mentioned the problem of failed hosts is handled differently. TCOT uses time-outs to abort transactions that can hold the resources for a long time. In pre-serialization, transaction of unresponsive nodes are not aborted immediately, rather they are aborted as new transactions need the resources held by those suspended. In pro-motion the data sent to mobile hosts are attached an expiration time so it is guaranteed that they will be released without blocking the other transactions. While using time-outs can be a problem because of the unpredictable transaction durations and messages exchanged for extension, it would be easier to predict if a node is dead for a long time or not.

None of these models tries to minimize the power used by mobile nodes. Moreover, in pro-motion mobile nodes execute transactions and exchange large amounts of data compared to the other ones. The other models try to minimize the messages passed on the wireless network, which in turn means mobile nodes will use less messages and power.

Models like Deno, clustered, and MANET are different from the models mentioned above in that they do not assume that there is a fixed network.

Deno is the most interesting model among all the models. It does not make any assumptions about the network topology, node capacities, and connection characteristics. It is a fully distributed model and it can handle network partitions seamlessly. However, it does not take advantage of the more powerful nodes if there are any in the network.

The MANET model assumes that there are more powerful nodes in the network. This is not an unreasonable assumption for the mobile networks. MANET can be used to take advantage of the wired network if there is any. It also tries to distribute the energy usage homogeneously, which is unique for all the models. Since this model assumes more powerful and always-connected nodes, disconnections can be handled similarly to the models mentioned above.

The clustered model is similar to Deno for that: it does not make assumptions about network topology. It also defines two types of transactions: weak and strong. Weak transaction can proceed when the network is partitioned.

The planned disconnected model supports transaction management in a disconnection mobile database, and increases flexibility and allows the distributed database to take advantage of mobility and use new ways.

The planned disconnected protocol produces executions that are one-copy serializable. Briefly, it can be argued that, since all of the transactions of the disconnected site are read-only, the values of the data that are read are those of a snapshot taken at the time of disconnection. All of the read-only transactions of the disconnected site can be serialized at the time of disconnection.

The check-out mode in a planned disconnected model with system reads produces executions that are one-copy serializable. The locked data items are modified by the transactions at the disconnected site, and these transactions are serialized with respect to each other because of local two-phase locking. They are serialized with respect to the transactions of the rest of the system at the point in time of reconnection (rather than the point of disconnection, as for check-out with mobile read).

SUMMARY AND CONCLUSION

The aim of this article was to present a comparison between a few of the transaction management models in distributed mobile databases. The study suggests that a few improvements still need to be accomplished in this context. Limitations such as low and inconsistent connections, low storage space, processing speed, and the fact that the computing is distributed made mobile databases subtle and prone to more difficulties.

The ACID model suggests that, in some situations, a client requests a sequence of separate requests to a server to be atomic, provided that they are free from interference by operations being performed on behalf of other concurrent clients; and either all of the operations must be completed successfully or have no effect at all in case of server crash.

Many of the transaction management models start from these concepts that ACID suggests. They normally use relaxing ACID as an architectural model. The kangaroo model, however, suggests that a new mobile transaction definition is needed, specific to the mobile computing proposed.

Each of the transaction management models makes its own assumptions about the infrastructure needed to support the respective model. For example the pre-serialization transaction management model and kangaroo model both give a very general architecture by which mobile computing can be performed in a heterogeneous multi-database environment.

The constantly decreasing price of mobile devices is leading to a revolution in the field of mobile computing. The set of enhanced services currently provided to the owners of PDAs is expected to reach the users of the next generation of mobile phones and other mobile devices. A set of very powerful applications is expected to support this revolution, ranging from simply text transfer up to even multimedia transfer.

REFERENCES

Cetintemel, U., & Kelender, P. (2002). Lightweight currency management mechanism in mobile and weakly-connected environment,. *Journal of Distributed and Parallel Database, 11*(January), 53-71.

Dircke, R., & Gruenwald, L. (2000). A Pre-Serialization transaction management technique for mobile multi-database. *ACM Mobile Networks and Applications, 5*(December), 311-321.

Dunham, M. H., & Helal, A. (1997). A mobile transaction model that captures both the data and the movement behavior. *ACM/Baltzer Journal on Special Topics in Mobile Networks and Applications, 2,* 149-162.

Gary, J., & Reuter, A., (1993). *Transaction processing: Concepts and technique.* San Francisco: Morgan Kaufman.

Gruenwald, L., & Banik, S. M. (2001, September). A power-aware technique to manage real-time database transactions in mobile ad-hoc networks. *Proceedings of the 4th International Workshop on Mobility in Database and Distributed*

Systems, part of the International Conference on Database and Expert systems Applications (DEXA).

Kumar, V., Prabhu, N., Dunham, M., & Seydim, Y.A. (2002). TCOT—A timeout-based mobile transaction commitment protocol. *IEEE Transactions on Computers, 51*(October).

Pitoura, E., & Bhargava, B. (1995). Maintaining consistency of data in mobile distributed environments. *Proceedings of the IEEE Workshop on Mobile Systems and Applications.*

Reihar, P., Heidemann, J. S., Ratner, D., Skinner, G., & Popek, G. J. (1994, June). Resolving file conflicts in the Ficus file system. *Proceedings of the USENIX Summer Conference* (pp. 183-195).

Skeen, D. (1985). Determining the last process to fail. *ACM Transactions on Computer Systems, 3*(1).

Unland, R., & Schlageter, G. (1992). A transaction manager development facility for non-standard database systems. In A. K. Elmagarmid (Ed.), *Database transaction models for advanced applications* (pp. 400-466). San Francisco: Morgan Kaufmann.

Walborn, G., & Chrysanthis, P. (1996). Transaction processing in promotion. *Proceedings of the ACM Symposium on Applied Computing.*

Ubiquitous and Pervasive Application Design

M. Bakhouya
The George Washington University, Washington DC, USA

J. Gaber
Université de Technologie de Belfort-Montbéliard, France

INTRODUCTION

The recent evolution of network connectivity from wired connection to wireless to mobile access together with their crossing has engendered their widespread use with new network-computing challenges. More precisely, network infrastructures are not only continuously growing, but their usage is also changing and they are now considered to be the foundation of other new technologies. A related research area concerns ubiquitous and pervasive computing systems and their applications. The design and development of ubiquitous and pervasive applications require new operational models that will permit an efficient use of resources and services, and a reduction of the need for the administration effort typical in client-server networks (Gaber, 2000, 2006). More precisely, in ubiquitous and pervasive computing, to be able to develop and implement applications, new ways and techniques for resource and service discovery and composition need to be developed. Service discovery is the process of locating which services are available to take part in a service composition. The service composition process so far concentrates on combining different available existing services as a result of the service discovery process. Most research to date in service discovery and composition is based on the traditional client/server interaction paradigm (CSP). This paradigm is impracticable in ubiquitous and pervasive environments and does not meet their related needs and requirements. Gaber (2000, 2006) has proposed two alternative paradigms to the traditional client/server interaction paradigm to design and implement ubiquitous and pervasive computing applications: the adaptive services/client paradigm (SCP) and the spontaneous service emergence paradigm (SEP).

Bio-inspired approaches are adequate to carry out these new paradigms for designing and implementing ubiquitous and pervasive applications (Gaber, 2000). Indeed, the adaptive servers/client paradigm, considered as the opposite of CSP, could be implemented via a self-adaptive and reactive middleware inspired by a biological system like the natural immune system. The service emergence paradigm could also be implemented by a natural system that involves self-organizing and emergence behaviors (Gaber, 2000).

Recently, agent-based approaches, with self-adapting and self-organizing capabilities, have been proposed in Bakhouya (2005) and Bakhouya and Gaber (2001, 2006a, 2006b) to implement SCP and SEP respectively. More precisely, these approaches, inspired by the human immune system, provide scalable and adaptive service discovery and composition systems for ubiquitous and pervasive environments.

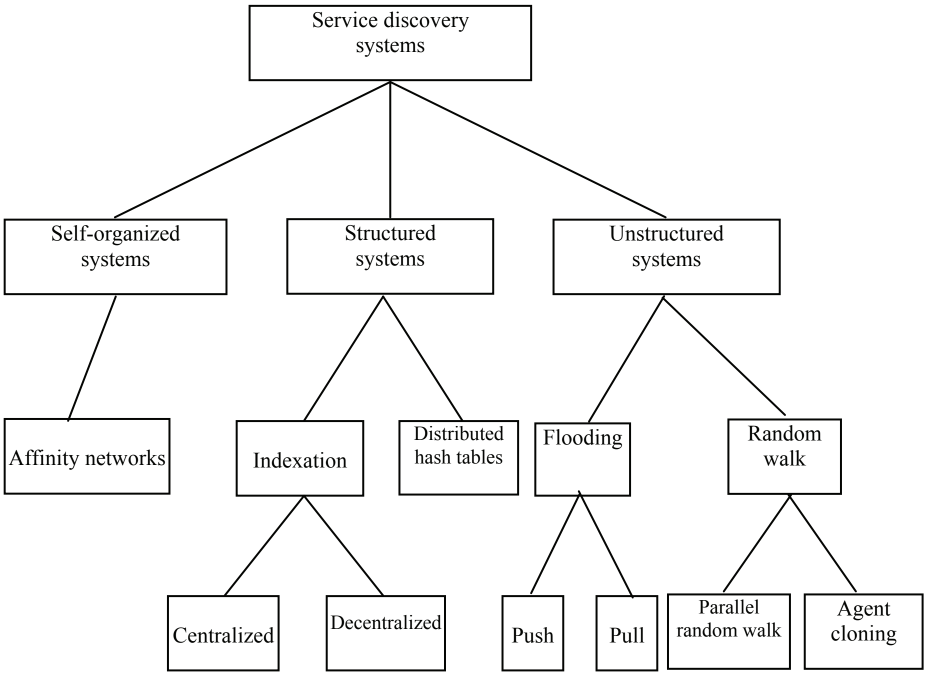
UBIQUITOUS COMPUTING

In ubiquitous computing (UC), the objective is to provide users the ability to access services and resources all the time and irrespective to their locations (Weiser, 1993). Service discovery and access systems can be classified into three categories as depicted in Figure 1: structured systems, unstructured systems, and self-organized systems. Structured systems can be classified also in indexation-based architectures and hashing-based architectures. In indexation-based architectures, there are two categories: centralized and decentralized systems. In centralized indexation-based systems, typical resource discovery architectures (Bettstetter & Renner, 2000), such as Jini (2001), consist of three entities: service providers that create and publish services, a broker that maintains a repository of published services to support their discovery, and services requesters that search the service broker's repository. Centralized approaches scale poorly and have a single point of failure. To overcome the scalability problem, decentralized approaches, such as m-SLP (Zhao, Schulzrinne, & Guttman, 2000) or Secure Service Discovery Service (Xu, Nahrstedt, & Wichadakul, 2001), traditionally have a hierarchical architecture consisting of multiple repositories that synchronize periodically. In hashing-based architectures (Wang & Li, 2003), proposed primary to file-sharing, distributed hash tables (DHTs) are used to assign files to specific nodes. This technique allows the implementation of direct search algorithm to efficiently locate files. However, hashing-based architectures require overlay networks between nodes that are generally hard to maintain.

In unstructured systems, the most typical localization mechanisms are flooding and random walk. There are two main flooding techniques: the push and the pull technique. In the first technique, the server advertises periodically its services across the network. The clients receive the service

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Classification of service discovery systems according to their architectures and their operating modes



advertisement and cache the information. This information must have a time period associated with it, and must be flushed out from the cache when this time period expires. Hence, the user has a complete knowledge of the available services, and no request resolution process is required. In the pull technique, the client has no knowledge of services present in the network. In this case, a service request is broadcast to all neighbors within a certain radius with a TTL (time to live) tag (Wang & Li, 2003). A random walk is a stochastic process that evolves in the following manner (Gaber & Bakhouya, 2006b). A client sends its query message (i.e., a walker) to a randomly chosen neighbor. At each step, the query message is forwarded to a neighbor of its current location, and the process continues this way by taking random steps that are independent of all the previous ones until meeting the required service. Consequently, the random walk technique avoids message duplication inherent to the flooding mechanism (Wang & Li, 2003). More precisely, by using one walker, it cuts down the message overhead significantly. Nevertheless, the delay for a successful request resolution could be high. To decrease this delay, a requester could send k parallel query messages, and each query message takes its own random walk. However, it is difficult to determine a priori a suitable value for k. In other words, if this number k is big enough, the message traffic could increase considerably. An alternative approach to avoid this problem uses both random walks together with an adaptive cloning agent-based technique for service discovery (Gaber & Bakhouya, 2006b).

It should be noted that the fundamental aspect of these systems is the process of service discovery based on the traditional client to server paradigm. More precisely, it is the user who should initiates a request, should know a priori that the required service exists, and should be able to provide the location of a server holding that service. This is why the use of repositories is essential in these discovery systems. However, ubiquitous environments have the potential ability to integrate a continuously increasing number of services and resources that can be nomadic mobiles and partially connected. A user can be mobile or partially connected, and its ability to use and access services will no longer be limited to those that she/he currently has at hand or those statically located on a set of hosts known a priori. Therefore, the ability to maintain, allocate, and access a variety of continuously increasing numbers of heterogeneous resources and services distributed over a mixed network (i.e., wired, wireless, and mobile network) is difficult to achieve with the traditional client/server approaches (Gaber, 2000, 2006). More precisely, these architectures cannot meet the requirements of scalability and adaptability simultaneously. The way in which they have typically been constructed is often very inflexible due to the risk of bottlenecks, the difficulty of repositories updating, or the network loading problem. This is particularly true for the cases where some services could be disconnected from the network and new ones may join it at any time.

An appropriate model was proposed originally by Gaber in Gaber (2000) as an alternative to the traditional client/

server paradigm. This model can be viewed as opposed to the client/server model and is denoted *adaptive servers/client paradigm.* In this model, it is the service that comes to the user. In other words, in this paradigm, a decentralized and self-organizing middleware should be able to provide services to users according to their availability and the network status. As pointed out in Gaber (2000), such a middleware can be inspired from biological systems like the natural immune system. The immune system has a set of organizing principles such as scalability, adaptability, and availability that are useful for developing a distributed networking model in a highly dynamic and instable setting. In Gaber (2000, 2006), the immune-based approach operates as follows: unlike the classical client/server approach, each user request is considered as an attack launched against the global network. The immune networking middleware reacts like an immune system against pathogens that have entered the body. It detects the infection (i.e., user request) and delivers a response to eliminate it (i.e., satisfy the user request).

Recently, an immune approach using mobile agents with cloning capabilities was proposed in Bakhouya (2005) and Bakhouya and Gaber (2006a, 2006b) to implement SCP. A mobile agent is a software program that may move from location to location to meet other agents or to access resources provided at each location. Using a mobile agent that can clone itself in order to increase system robustness and performance is an attractive idea. The clone operation creates multiple instances of an agent that runs on different machines. For example, an initial mobile agent starts on the requesting machine and, after a local step, creates replications (i.e., clones) that initiate parallel walks to further machines. This would allow agents to cover a much wider area of machine space in a reasonable amount of time (Gaber & Bakhouya, 2006b). However, it should be noted that increasing agent population with cloning operation will increase resource demands in the network, which would indirectly affect network performance. Since mobile agents operate in a dynamic and distributed environment, it is difficult, even impossible, to estimate a priori an appropriate number of agents in the network (Bakhouya, 2005). Also, changing the population dynamically in response to its environment is a complex issue in the absence of a central controller. A distributed approach for the regulation of mobile agent population and inspired by the immune system is proposed in Bakhouya (2005) and Bakhouya and Gaber (2002, 2006b).

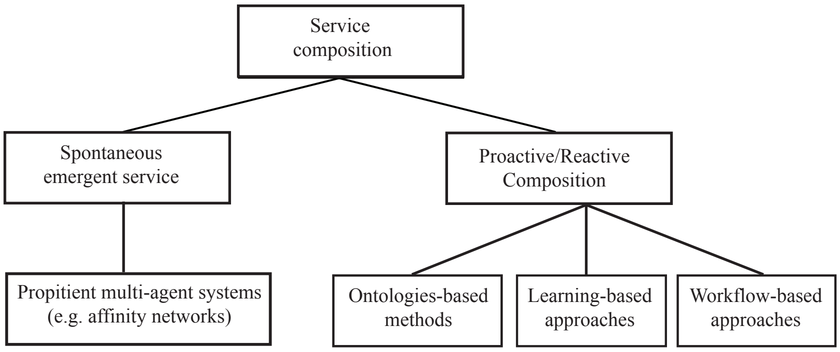
The immune system consists primarily of lymphocytes that circulate through the body in the blood and lymph system. There are two categories of lymphocytes, the B-cells and T-cells. The B-cells are developed in the bone marrow and the T-cells are developed in the thymus. The principle function of T-cells is to potentate the immune response by the secretion of specialized factors that activate other cells to fight off infection. The major function of the B-cell is the production of antibodies in response to foreign antigens. According to Jerne, B-cells are interconnected by affinity relationship against foreign antigens and form idiotypic networks (Jerne, 1974).

The mapping between the immune system entities and the middleware agents is done in the following manner. T-cells represent servers while B-cells represent services and resources. Antigens correspond to client requests while antibodies correspond to delivered responses. Servers are organized into communities by the creation of affinity relationships in order to represent services in the network. The establishment of relationship affinities between servers allows the solving, by collaboration, of user requests. A reinforcement learning mechanism is used to adjust and reinforce dynamical relationship affinity values according to delivered responses (Bakhouya, 2005). This reinforcement mechanism permits coping with dynamic changes in the network, the services availability, and the user requests. Similar to the natural immune system, new communities may be created or modified according to a dynamically changing environment. In other words, servers may acquire new or drop current servers through establishing or deleting the affinity relationship.

PERVASIVE COMPUTING

Pervasive computing (PC), often considered the same as ubiquitous computing in the literature, is a related concept that can be distinguished from ubiquitous computing in terms of environment conditions. We can consider that the aim in UC is to provide any mobile device access to available services in an existing network all the time and everywhere, while the main objective in PC is to provide spontaneous services created on the fly by mobiles that interact by ad hoc connections (Gaber, 2000, 2006). Service composition systems can be classified into three categories as depicted in Figure 2: proactive composition systems, reactive composition systems, and spontaneous emergent service systems. The first category refers to off-line composition of available services to form new ones. Services that may be used for proactive service composition can be considered stable, widely and always available (Chakraborty & Joshi, 2001). The second category refers to the process of creating a composite service on demand. In other words, a composite service is created only when a user requests the execution of that service. Most known reactive and proactive service composition systems, such as the eFlow system (Casati, Ilnicki, Jin, Krishnamoorthy, & Shan, 2000), are based on a centralized broker which manages the service composition process (Chakraborty & Joshi, 2001). The drawback is that if a huge number of users attempt to access a variety and increasing number of services distributed over the network, the broker quickly becomes a bottleneck. It should be noted also that these systems are based on the client/server paradigm; it is the user who should

Figure 2. Classification of service composition systems according to their architectures and their operating modes


Service composition
Spontaneous emergent service
Proactive/Reactive Composition
Propitient multi-agent systems (e.g. affinity networks)
Ontologies-based methods
Learning-based approaches
Workflow-based approaches

initiate a request, and moreover, services and future demands are known in advance.

Gaber (2000, 2006) has proposed a second alternative paradigm to the client/server one for service composition that suits pervasive environments. This paradigm involves the concept of spontaneous emergence and is called the *spontaneous service emergence paradigm.* This paradigm can also be carried out by an inspired natural immune middleware that allows the emergence of ad hoc services on the fly according to dynamically changing context environments such as computing context and user context (Gaber, 2000). More precisely, in this model, ad hoc or composite services are represented by an organization or group of autonomous agents. Agents correspond to the immune system B-cells. Agents establish relationships based on affinities to form groups or communities of agents in order to provide composite services. A community of agents corresponds to the idiotypic network in the human immune system (Gaber, 2000).

More generally, agents together with their affinity relationships as a whole form a *propitient multi-agent system* (Bakhouya & Gaber, 2006c). A propitient system is a system with the ability to self-organize in order to adapt towards the most appropriate agent organization structures according to unpredictable changes in the environment. This emergent behavior is delivered as a result of agent-to-agent and agent-to-environment interactions that adapt until the system hits a most suitable affinity network. Therefore, a propitient multi-agent system implements the SEP.

A self-organizing approach assumes that individual agents are autonomous agents, while multi-agent organizations are emerged structures that are not represented explicitly, but they exist through the affinity relationships between agents. In other words, agents cooperate equally rather than being assigned subordinate and supervisory relationships. It is worth noting that this multi-organization based on dynamic affinities supported by relationships provides a highly decentralized system while remaining adaptive in dynamic and open environments. More precisely, this decentralized organizational structure offers a high degree of resilience against an agent leaving the organization. For example, when an agent leaves an organization, all the peer affinity relationships with other agents are removed without additional messages since it does not rely on any overlay control structure. An affinity-driven clustering learning mechanism could be used to adjust the affinity relationships between nodes to cope with the user context and provoke or produce an emergent service (Gaber, 2000; Gaber & Bakhouya, 2006a). More precisely, an ad hoc emergent service is created spontaneously on the fly for a user or between a group of users in an unpredictable manner (i.e., without a priori intention).

CONCLUSION

The design and development of ubiquitous and pervasive applications require alternative operational models to the traditional client/server paradigm. The adaptive servers/client paradigm and spontaneous service emergence paradigm are more adequate to ubiquitous and pervasive computing respectively. Service discovery and composition systems based on these three paradigms and proposed in the literature are presented with emphasis on elf-organizing and self-adapting approaches inspired by the immune system to implement SCP and SEP. Self-adaptation and self-organization are crucial issues in systems that operate in an open and dynamic environment.

REFERENCES

Bakhouya, M. (2005). *Self-adaptive approach based on mobile agent and inspired by human immune system for service discovery in large scale networks.* PhD thesis NO. 34, Université de Technologie de Belfort-Montbéliard, France.

Bakhouya, M., & Gaber, J. (2002) Distributed autoregulation approach of a mobile agent population in a network (Research Report RR-12-02, pp. 1-14). Université de Technologies de Belfort-Montbéliard, France.

Bakhouya, M., & Gaber, J. (2006a). Adaptive approaches for ubiquitous computing. *Mobile networks and wireless sensor networks* (pp. 129-163). Hermes Science.

Bakhouya, M., & Gaber, J. (2006b). Adaptive approach for the regulation of a mobile agent population in a distributed network. *Proceedings of the 5th International Symposium on Parallel and Distributed Computing (ISPDC'06)* (pp. 360-366). Timisoara, Romania: IEEE Press.

Bakhouya, M., & Gaber, J. (2006c). Self-organizing approach for emergent multi-agent structures. *Proceedings of the Workshop on Complexity Through Development and Self-Organizing Representations* (GECCO'06) (pp. 1-5). Seattle, WA: ACM Press.

Bettstetter, C., & Renner, C. (2000). A comparison of service discovery protocols and implementation of the service location protocol. *Proceedings of the 6th EUNICE Open European Summer School.* Retrieved December 19, 2006, from http://www.bettstetter.com/publications/bettstetter-2000-eunice-slp.pdf

Casati, F., Ilnicki, S., Jin, L., Krishnamoorthy, V., & Shan, M. (2000). *Adaptive and dynamic service composition in eFlow.* Technical Report HPL-200039, Software Technology Laboratory, Palo Alto, CA.

Chakraborty, D., & Joshi, A. (2001). *Dynamic service composition: State-of-the-art and research directions.* Technical Report TR-CS-01-19, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, USA. Retrieved from http://citeseer.ist.psu.edu/chakraborty01dynamic.html

Gaber, J. (2000). *New paradigms for ubiquitous and pervasive computing* (Research Report RR-09). Université de Technologies de Belfort-Montbéliard, France.

Gaber, J. (2006). New paradigms for ubiquitous and pervasive applications. *Proceedings of the 1st Workshop on Software Engineering Challenges for Ubiquitous Computing,* Lancaster, UK.

Gaber, J., & Bakhouya, M. (2001), A middleware for large scale networks inspired by the immune system (Research Report RR-11-01, pp. 1-6). Université de Technologies de Belfort-Montbéliard, France.

Gaber, J., & Bakhouya, M. (2006a). An affinity-driven clustering approach for service discovery and composition for pervasive computing. *Proceedings of the IEEE International Conference on Pervasive Services (ICPS'06)* (pp. 277-280). Lyon, France.

Gaber, J., & Bakhouya, M. (2006b). Mobile agent-based approach for resource discovery in peer-to-peer networks. *Proceedings of the 5th International Workshop on Agents and Peer-to-Peer Computing* (at AAMAS) (pp. 1-9). Hakodate, Japan.

Hofmeyr, S. A., & Forrest, S. (2000). Architecture for an artificial immune system. *Evolutionary Computation, 8*(4), 443-473.

Jerne, N. (1974). Towards a network theory of the immune system. *Annals of Immunology, 125,* 125-373.

Jini. (2001). *Jini technology core platform specification.* Retrieved from http://www.sun.com/jini/specs

Robert, M. (2000). *Discovery and its discontents: Discovery protocols for ubiquitous computing.* Research Report UIUCDCS-R-99-2132, Department of Computer Science, University of Illinois Urbana-Champaign, USA.

Wang, C., & Li, B. (2003). *Peer-to-peer overlay networks: A survey.* Technical Report, Department of Computer Science, HKUST. Retrieved from http://comp.uark.edu/cgwang/

Watanabe, Y., Ishiguro, A., & Uchkawa, Y. (1999). Decentralized behavior arbitration mechanism for autonomous mobile robot using immune system. *Books artificial immune systems and their applications.* Berlin: Springer-Verlag.

Weiser, M. (1993). Hot topics: Ubiquitous computing. *IEEE Computer.*

Xu, D., Nahrstedt, K., & Wichadakul, D. (2001). QoS-aware discovery of wide-area distributed services. *Proceedings of the 1st IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid)* (pp. 92-99). Brisbane, Australia.

Zhao, W., Schulzrinne, H., & Guttman, E. (2000). mSLP-mesh-enhanced service location protocol. *Proceedings of the International Conference on Computer Communications and Networks* (ICCCN 2000) (pp. 504-509). Retrieved from draft-zhao-slp-da-interaction-07.txt

KEY TERMS

Adaptive Services/Client Interaction Paradigm (SCP): Adaptive Services to Client interaction paradigm is the opposed model to the traditional Client/Server interaction paradigm in which it is the most appropriate service that comes to the user in response to a request. This most suitable service can be computed by the network itself via an intelligent middleware.

Mobile Agents: A mobile agent is a software entity which may move with its own code and execution context from node to node to meet other agents or to access resources provided at each location.

Pervasive Computing: The main objective of pervasive computing is to provide spontaneous emergent services created on the fly by mobiles that interact by ad hoc connections

Propitient System: An agent-based system with the ability to self-organize in order to adapt towards the most appropriate organization that copes with the unpredictable changes in the environment.

Reinforcement Learning: A mechanism used to adjust and reinforce dynamically relationship affinity values according to delivered responses and in order to cope with dynamic changes in the network, the service's availability and user requests.

Service Composition: Service composition process concentrates on combining different available and selected services via the service discovery process to deliver new ones.

Service Discovery: Service discovery is the process of locating which services are available to take place in a service composition.

Spontaneous Service Emergence Paradigm (SEP): Allows the emergence of ad hoc services on the fly for a user or group of users without a priori planning or intention

Traditional Client/Server Interaction Paradigm (CSP): The traditional Client to Server paradigm is based on the following fundamental aspect: it is the user who should initiate a request, should know a priori that the required service exists and should be able to provide the location of a server holding that service. It should be noted that push, pull and P2P systems are still based on the Client/Server interaction paradigm.

Ubiquitous Computing: the main objective of ubiquitous computing is to provide users with the ability to access services and resources all the time and irrespective to their location.

The "Umbrella" Distributed Hash Table Protocol for Content Distribution

Athanasios-Dimitrios Sotiriou
National Technical University of Athens, Greece

Panagiotis Kalliaras
National Technical University of Athens, Greece

INTRODUCTION

During the past few decades, the Internet has blossomed due to the immerse growth of the telecommunication backbone, making it one of the key players in a wide area of fields. Even traditional players such as television or radio are now being challenged by the new entertainment media, the home computer. The increase of share communities such as Weblogs (Drezner & Farrell, 2004) or MySpace (www.myspace.com) and content-sharing software proves that people want to share their content with their global Web community. The need for such content-sharing software is therefore undisputable. Such attempts have been introduced in many ways during the past, with perhaps the most common example being Napster (www.napster.com) and Gnutella (http://gnutella.wego.com).

The technical and ethical issues of these systems proved to be their weak point. Systems that have no central point of control and distribute functions among all users seem better fit for sharing and distributing content. A solution has been proposed in the form of *distributed hash-tables (DHTs).* This article proposes an alternative architecture for content distribution based on a new DHT routing scheme. The proposed architecture is well structured and self-organized in such a way as to be fault-tolerant and highly efficient. It provides users with content distribution and discovery capabilities on top of an overlay network. The novelty of our proposed architecture lies in its routing table which is maintained by each node and is of constant size, as opposed to other algorithms that are proportional to the network's size (usually O(logN)). All operations in our architecture are of O($\log_b$N) steps (entry, publishing, and lookups) and degrade gracefully as up-to-date information of the routing table decreases due to numerous node failures.

BACKGROUND

The firsts to introduce routing algorithms that could be applied to DHT systems were Plaxton, Rajaraman, and Richa (1997). The algorithm was not developed for P2P systems, and thus every node had a neighborhood of O(logN) and inquires resulted in O(logN) steps. It was based on the ground rule of comparing one byte at a time until all bytes of the identifier (or best compromise) were met. Our scheme meets the logarithmic growth of inquiries introduced by Plaxton et al. (1997), even though nodes are not placed within constant distance from each other.

A variation of the Plaxton algorithm was developed by Tapestry (Zhao et al., 2004), properly adjusted for P2P systems (where overall state is not available). The algorithm once again tackles one digit at a time, and through a routing table of $\beta*\log_\beta$N neighbors routes to the appropriate node, resulting in a search of $\log_\beta$N maximum steps.

Pastry (Rowstron, 2001) is similar to Tapestry, but added a leaf set of neighbors that the node first checks before referring to the routing table. Also a different neighbor set is maintained for tolerability issues. Each node maintains a neighborhood of $\log_2$bN rows with (2^b-1) elements in each row and requires a maximum of O($\log_2$bN) steps for enquires. Proper routing is maintained as long as (L/2) nodes are available in the neighborhood of each node. Once again, the variable size of each node's table (which must be maintained up-to-date) limits the algorithm's scalability.

In Chord (Stoica, Karger, Kaashoek, & Balakrishnan, 2001) a different approach was applied, placing nodes in a circular space and maintaining information only for a number of successor and predecessor nodes through a finger table. Routing is established through forwarding queries to the correct successor based on the identifier. Even though the basic Chord mechanism only requires the knowledge of one successor, modifications where needed in order for the system to be applicable to a robust environment, introducing a finger table of O(logN) size.

Finally, Kademlia (Maymounkov & Mazieres, 2002) bases nodes in a binary-tree through identifiers. Each node of the tree retains information concerning one node from each leaf, other than the one in which it resides. It also differentiates by applying an XOR comparison on identifiers instead of the casual comparison of each bit, adopted by all other algorithms.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

SYSTEM ARCHITECTURE

In this section we will give an overview of the Umbrella architecture. The main functions consist of the insertion of nodes, the assignment of keys to corresponding nodes, and the routing mechanisms for three principle operations, namely the insertion of nodes, publication of content, and lookup of keys. Our architecture is based on an overlay network, and thus we assume that node connectivity is both symmetric and transitive.

Hashing Function

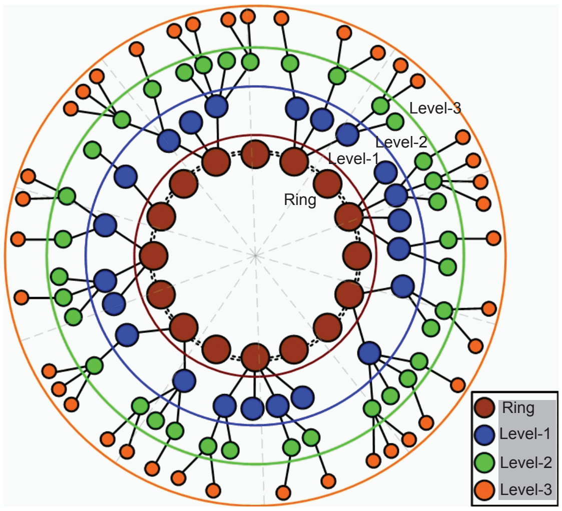
The proposed architecture is based on the creation of an overlay network, where all inserting nodes are identified by a unique code, asserted by applying the SHA-1 (NIST, 1995) hash-function on the combination of IP and computer name, which returns a 160-bit identifier. This hash-function has been proven to distribute keys uniformly in the 160-bit space and thus provides the desired load balancing for both the user space and the content space, as the same function is applied to each content destined for distribution in the system.

Structure Overview

The main objective of the architecture is to insert and retain nodes in a simple and well-structured manner, thus querying and fetching of content is both efficient and fault-tolerant. In addition, each node will need only to retain up-to-date information of a limited, constant number of neighboring nodes, allowing the system to escalate in population of both users and content. Each node is inserted into the system through an existing node, which announces the new entrance. When this procedure has ended successfully, the new node can, having acquired and informed all neighboring nodes, continue to publish all of its content. The publishing procedure is similar to the insertion mechanism, as content is characterized by a number of keys, which after being hashed can be forwarded in the same manner. All keys are published in an existing node whose identifier is the closest match to the key identifier. In a similar fashion, querying is performed by routing the request to the node with the identifier closest to the desired key. If no such node exists, it is assumed that the desired content is not available.

The overlay network is constructed in the form of a loose B-Tree, where each node is placed in a hierarchy tree with a parent node and b child nodes, which in our initial architecture is of the value 16 (in order to classify the 160-bit identifiers to a maximum B-Tree of height ≤ 40). All nodes are placed along the tree structure, without being required to fulfill pre-defined ranges as in a proper B-tree structure, and are responsible for updating their connections with neighboring nodes that reside on either the parent, sibling, or child level. Along with obvious connections (parent, child, and sibling level of each node), further links to a limited number of nodes in the near vicinity are kept in record for fault-tolerant operations. Figure 1 illustrates the structure of this loose B-Tree. Routing in the umbrella protocol is simple and constitutes the forwarding of messages to either a parent or child node until the appropriate node is reached. In the rest of the article, with the term-appropriate node we will refer to either the exact or closest match alike.

Figure 1. The Umbrella architecture



Key Mapping

Each level n of the structure is capable of withholding b^{n+1}nodes. Each node has a unique parent node, which is always one level higher, and a maximum of b children at a lower level. The Umbrella overlay network is configured with the following simple rule. The relation between a parent node at level n and a child node (which must by default reside on level n+1) is defined as such and only such that:

- The n+1 first (from left to right) digits of the parent's identifier are equal to the corresponding digits of the child's identifier.
- The n+2 digit of the child's identifier determines the child's position in the parent's child list. Thus all children of the same parent share the first n+1 digits and all differ in the n+2 digit.

The above simple rule is obeyed by all nodes entering the Umbrella overlay network, with only the exception of the first node that actually initiates the network and is considered to be positioned at level -1. As already stated, the SHA-1 hash

Table 1. Fields of the neighborhood table

Field	Set	Description
Level	Basic	The level it resides
Right	Basic	The non-empty node to the right
Left	Basic	The non-empty node to the left
Up	Basic	The parent node
Right2	Upper	The node residing to the right of the parent node
Left2	Upper	The node residing to the left of the parent node
Up2	Upper	The parent’s parent node
Right3	Lower	One (random) child of the node to the right
Left3	Lower	One (random) child of the node to the left
Umbrella	Basic	All child nodes
Umbrella2	Lower	A (random) child node from each child

function is used to assign identifiers to both nodes and content, offering a uniform distribution in the 160-bit space along with non-voluntary placement anonymity (Milojicic et al., 2002) of the published content. In order to apply the routing algorithms, we define a comparing function for identifiers as *comp,* which compares two identifiers and calculates their difference as a long integer, with importance given to digits from left to right. The consistent hash function balances key distribution among nodes, as stated in Karger et al. (1997) in the form of the following theorem:

Theorem 1. Given a set of nodes N and keys K, then with high probability each node is responsible for an average of K/N keys, with a maximum of (1+α)K/N, whereas α is a parameter with bound of O(logN).

Routing Table

As in most DHT systems, a routing table is maintained by each node in order to route incoming messages. Each node is responsible for keeping the table up-to-date by issuing messages to all nodes in its table at different intervals. The routing table in our architecture consists of three different sets—a basic, an upper, and a lower set. The basic set stores nodes and information needed for basic routing operations under fault-free conditions. The upper and lower sets store additional indexes to nodes in the upper and lower levels, correspondingly, which are utilized when nodes in the basic set become unreachable. These three sets constitute the node’s neighborhood table and are presented in Table 1.

The above elements are sufficient to maintain proper routing in our architecture even in the case of sudden failure of nodes. The upper set allows routing to nodes of higher level (when the parent node is unreachable) and the lower set to nodes of lower level (when child nodes fail). Each node is responsible to modify or fix its routing table when nodes enter/leave the network or a failure to communicate with another node is detected, respectively. Our architecture’s structure and routing table described so far ensure that a published key can be located by an appropriate query within logarithmic overlay steps to the total size of the network. This is stated and proved within the following two theorems:

Theorem 2. Given an Umbrella network of N nodes with identifiers of base b acquired by a consistent hash function, the maximum height of the loose B-tree structure is of logarithmic scale.

Proof: Let b denote the base of our identifiers, N the total number of nodes, and k a particular level in the Umbrella structure. Then according to the Umbrella protocol, in each level a maximum of b^k nodes can reside, with $b^0=1$ as stated for the first node that creates the network. Thus, if m denotes the number of levels required for the above population of nodes, we acquire the following relation, in respect to Theorem 1 that provides, with high probability, a uniform distribution of identifiers to our space:

$$N = \sum_{k=0}^{m} b^k = \frac{b^0 - b^{m+1}}{1-b} = \frac{1-b^{m+1}}{1-b} \Leftrightarrow$$

$$\Leftrightarrow m = \{\log_b [N(b-1)+1]\} - 1$$

Thus the maximum height m of our structure is of $O(\log_b N)$.

Theorem 3. A successful lookup in an Umbrella network requires, with high probability, $O(\log_b N)$ steps.

Proof: Suppose that a node p that resides at level l_p is seeking a specific key k that resides within our network in another node f at level l_f. If m denotes the number of levels of the current network, N the nodes, and b the base of identifiers, then we could argue that the worst-case scenario would require both nodes to reside at level m and with maximum distance between them (thus node p is an m-depth child of the first child at level 0 and on-forth, and node f is the m-depth child of the b child at level 0 and on-forth). In this case, the lookup must first ascend all the way to the top of our structure (thus m steps) and then descend to the bottom (m steps again). In total, a maximum of $2m$ steps are required. Hence, from Theorem 2, the required maximum steps for a successful lookup is, with high probability, of $O(\log_b N)$ steps.

ALGORITHMS AND IMPROVEMENTS

Main Algorithms

During the creation of the overlay network, the first node to enter creates the new network by placing itself on the top of the system. As new nodes arrive, they are placed according to their identifier, as described in the previous section. A node only needs to contact an existing node in the system in order to be inserted (special mechanisms for fetching existing nodes by outside contacts are not of the scope of this article, as our architecture can embody any of numerous such techniques already proposed (Francis, 1999)). Only the first node is automatically inserted regardless of its identifier; all subsequent nodes are placed within the system based on the insertion algorithm. The insertion mechanism is quite simple, intentionally, as with all of the system's mechanisms, and consists of the following steps:

- Contact an already connected node and issue a request for insertion.
- The established node checks if the n+1 first digits of the identifier match its own, where n is the level the node resides.
- If not, the insertion message is forwarded to the node's parent.
- If yes, the message is forwarded to the child with the n+2 digit common with that of the new node.
- If such a child does not exist, then the new node is placed as a child to the current node and is informed of its new neighbors and via versa.

The publish procedure is similar to insertion and is therefore suppressed. Conversely, the search mechanism is executed as follows:

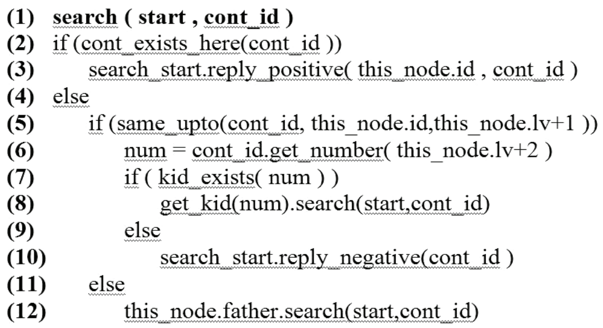
- The node first checks for the keyword in its list of published keywords.
- If it exists then the search terminates.
- If not, then it checks whether the first n+1 digits are identical to its own identifier.
- If not, the message is forwarded to its parent.
- If yes, then it is forwarded to the child with the corresponding n+2 digit matching.
- If no such child exists, then the search fails.

The pseudo code of the above algorithm is given in Figure 2.

The final mechanism provided by our protocol is that of node departure from the system. When a node issues a departure, the following steps are followed:

- If the node has no children, then all of its keywords are forwarded to its parent and it informs all its neighbors of its departure.
- If it has any child, then it randomly picks one and copies all of its neighborhood and keyword information to it before departing. The chosen child moves up a level and substitutes the departing node.
- If the chosen child has any child, then the previous step is repeated recursively until a node with no children is reached and the first step is then executed ending the algorithm.

Figure 2. Pseudo code for search mechanism



```
search ( start , cont_id )
if (cont_exists_here(cont_id ))
    search_start.reply_positive( this_node.id , cont_id )
else
    if (same_upto(cont_id, this_node.id,this_node.lv+1 ))
        num = cont_id.get_number( this_node.lv+2 )
        if ( kid_exists( num ) )
            get_kid(num).search(start,cont_id)
        else
            search_start.reply_negative(cont_id )
    else
        this_node.father.search(start,cont_id)
```

Enhanced Algorithms

The algorithms presented in the previous section embody the main mechanisms of our architecture and are capable of maintaining the system stable and fully functional under normal conditions, as will be validated by our simulation results in the following article. The system is however liable to node departures, either intentional or due to network disconnections. In the next set of mechanisms, we will concentrate on sudden departures of nodes, which we will call "node failures." These are due to either voluntary departure without calling the appropriate mechanism, sudden departures due to client errors, or nodes becoming unavailable due to network disconnections. We will treat all of the above cases in the same manner, and through changes in the algorithms already presented, we will allow the system to bypass node failures. Most changes are based on using the upper and lower set of our neighborhood table to bypass nodes that are not responding. The upper set is utilized to forward messages to nodes of a higher level, while the lower set is for nodes on a lower level.

In the first case, when a node is unable to contact its parent node, it attempts to forward requests consequently to:

- the parent's parent node (field Up2 on the upper set),
- the node to the right of the parent node (field Right2 on the upper set), and
- the node to the left of the parent node (field Left2 on the upper set).

Whichever of the above succeeds first will terminate the mechanism. The only exception to the above algorithm is that of the node on level -1, which triggers a variation since it has no parent of sibling nodes, and attempts are made toward nodes on either the left or the right of the child node.

In the latter case of a child node failure, the corresponding nodes are contacted in the following order:

Table 2. Corresponding forwarding of repair messages in order to reach a parent node

Failing Node	Action Taken
Up	Contact Up2
Node2	Contact Node
Left3 or Right3	Contact Left or Right respectively
Left2 or Right2	Contact Up2
Left or Right	Contact Up
Up2	Contact Up

- one of the child’s child (field Umbrella2 on the lower set),
- the node on the right of the child (field Umbrella on the basic set),
- the node on the left of the child (field Umbrella on the basic set),
- a child of the node right of the issuing node (field Right3), and
- a child of the node left of the issuing node (field Left3).

Repair Mechanism

In order to address the problem of node failures even further, we have designed a repair mechanism, which is invoked whenever such a failure is detected. The algorithm utilizes the delete algorithm presented in our main mechanisms section in order to repair a failure to a child node. It can be proven that all other failures can be transformed into a child failure through contacting nodes in the neighboring table and forwarding a repair message to higher or lower levels until the parent node of the failing node is reached. More details of the actions involved in order to resolve the appropriate node are given in Table 2.

Once the appropriate node is reached and informed of the child failure, a variation of the delete algorithm is evoked in order to repair the failure by substituting the failed node with one of its children or by deleting it if none is available. Each node is responsible for checking its neighborhood table periodically by issuing ping messages to all node entries and invoking the repair mechanism whenever a failure is detected.

SYSTEM EXTENSIONS

Having presented the core structure and logic behind our routing protocol, we will continue with a number of extensions that improve the system’s performance. The first extension introduces the use of replication schemas, which have been shown to increase the robustness of content distribution systems (Ghodsi, Alima, & Haridi,, 2005). In this article we have implemented three additional replication schemas.

Table 3. Singular identifier assignment functions

Name	Details	Instance
OI	This is the base fuction	a b c d e fg h
II	This function inverses the identifier	h g f e d c b a
IP	The identifier’s digits are inversed by pair	b a d c f e h g
IPW	All digits are inversed by pair as in the case of IP and the result is inversed as a whole as in II	g h e f c d a b
IH	The II function is applied to the first and second half of the indentifier indenpendently	d c b a h g f e
SH	The first and second halves are switched without being inverted	e f g h a b c d
RR	A random reordering of the identifier’s digits	d a e c g h b f
SRR	Same as the case of RR but with different random generator	c f b a h g d e

Our core routing protocol publishes a keyword in a single node, the one with the closest identifier to that of the keyword. All three replications schemas retain this quality and enhance it by also publishing the keyword to a number of additional nodes, from which one can recall a successful lookup. Our three variations follow:

1. **Local Spread Replication (LSR):** The keyword is also published in all nodes residing in its neighboring table.
2. **Inverse Replication (IR):** This mechanism publishes keywords to the closest match and to the inverse closest match.
3. **Local Spread Inverse Replication (LSIR):** It implements a local spread in both the closest and the inverse closest match.

The second extension implemented allows nodes to participate in a number of virtual networks, with a different identifier in each one. This allows each node to have a different set of neighbors and thus increase its tolerability substantially. In order to achieve this, we have defined a number of singular identifier assignment functions that transform the original identifiers into a new set of identifiers. This new set is then used to allocate nodes and route requests in the virtual networks. We have defined seven different such functions, which are given in Table 3.

CONCLUSION

Through the course of this article, we presented the Umbrella protocol, a novel protocol based on a distributed hash table that supports key publishing and retrieval on top of an overlay network for content distribution. We have analyzed our protocol and its algorithms through theoretical means, and provided a number of algorithms and extensions. Its main novelty lies in its fixed-size routing table sustained by each node, which is able to provide efficient routing even under contrary conditions. The protocol is also highly scalable due to its low traffic load demands.

REFERENCES

Drezner, D., & Farrell, H. (2004). Web of influence. *Foreign Policy Magazine.*

Francis, P. (1999). *Yoid: Extending the multicast Internet architecture.* White Paper. Retrieved from http://www.aciri.org/yoid

Ghodsi, A., Alima, L.O., & Haridi, S. (2005). Symmetric replication for structured peer-to-peer systems. *Proceedings of the 3rd International Workshop on Databases, Information Systems and Peer-to-Peer Computing,* Trondheim, Norway.

Karger, D., Lehman, E., Leighton, F., Levine, M., Lewin, D., & Panigrahy, R. (1997). Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web. *Proceedings of the 29th Annual ACM Symposium on Theory of Computing,* El Paso, TX.

Maymounkov, P., & Mazieres, D. (2002). Kademlia: A peer-to-peer informatic system based on the XOR metric. *Proceedings of IPTPS'02,* Cambridge, MA.

Milojicic, D., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., et al. (2002). *Peer-to-peer computing.* HPL-2002-57R1, HP Labs Technical Report.

NIST (National Institute of Standards and Technology). (1995). *FIPS Pub 180-1: Secure Hash Standard (SHA-1).* Federal Information Processing Standards Publication.

Plaxton, G., Rajaraman, R., & Richa, A. W. (1997). Accessing nearby copies of replicated objects in a distributed environment. *Proceedings of the 9th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*.

Rowstron, D.P. (2001). Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. *Proceedings of Middleware 2001.*

Stoica, M. R., Karger, D., Kaashoek, F., & Balakrishnan, H. (2001). Chord: A peer-to-peer lookup service for Internet applications. *Proceedings of SIGCOMM.*

Zhao, B. Y., Huang, L., Stribling, J., Rhea, S. C., Joseph, A. D., & Kubiatowicz, J. D. (2004). Tapestry: A resilient global-scale overlay for service deployment. *IEEE Journal on Selected Areas in Communications.*

KEY TERMS

DHT: Distributed hash table.

Hashing: Producing hash values for accessing data or for security. A hash value (or simply hash), also called a message digest, is a number generated from a string of text. The hash is substantially smaller than the text itself, and is generated by a formula in such a way that it is extremely unlikely that some other text will produce the same hash value.

Overlay Network: A network built on top of one or more existing networks. This network adds an additional layer of indirection/virtualization and also changes properties in one or more areas of underlying network.

P2P: Peer-to-peer.

Replication: A duplicate copy of similar data on the same or a different platform.

Routing: The process of moving a packet of data from source to destination. Routing in P2P networks refers to finding a path to the desired node.

SHA-1: Secure hash standard.

Understanding Multi-Layer Mobility

Sasu Tarkoma
Helsinki Institute for Information Technology, Finland

Jouni Korhonen
TeliaSonera Corporation, Finland

INTRODUCTION

Mobility is an important requirement for many application domains, where entities change their physical or logical location. Physical location denotes the real-world location of a device, whereas logical location is not necessarily dependent on the physical environment. Mobility support may be divided into several technical layers and also categories depending on the nature of mobility. In this article, we consider mobility protocols starting from the network layer (layer 3 in the OSI stack) and ending at the application layer (layer 7), and focus on physical mobility.

The most fundamental network-level protocols for supporting mobile hosts are the Mobile-IP protocols standardized by the IETF (Perkins, 2002; Johnson, Perkins, & Arkko, 2004). Another related network-level solution is *network mobility* (NEMO) (Devarapalli et al., 2004), in which complete sub-networks may change location as well as single hosts. Mobility can also be handled on the transport layer. *Transport layer seamless handover* (TraSH) (Fu et al., 2004), *datagram congestion control protocol* (DCCP) (Kohler, 2006), and mSCTP (Xing, Karl, Wolisz, & Müller, 2002) are recent examples of such solutions. Yet another way of managing host mobility is with mobility-aware *virtual private networks* (VPNs) such as MOBIKE-based IPSec VPNs (Kivinen, 2006). Protocols such as wireless CORBA (WCORBA) (OMG, 2004) and the *session initiation protocol* (SIP) (Schulzrinne & Wedlund, 2000) provide more fine-grained mobility than host based, and they do not assume underlying transport- or network-level mobility support.

Middleware support for mobility is required in order to provide location transparency for objects, agents, and other components; support efficient and reliable communication in wireless environments; and buffer messages and other data for disconnected operation. In addition, the middleware may support scalability and availability of resources and services.

Mobility is inherently tied with the way nodes are addressed in a distributed network. In this article, we examine three different ways to address mobile nodes and components: addresses with *location and identity, locator/identity split,* and *content-based addressing.* The first addressing model is used by the IP protocol. The second model is an extension of the first and used, for example, in the *host identity protocol* (HIP) and the i3 overlay (Stoica, Adkins, Zhuang, Shenker, & Surana, 2002). The third model has been proposed for expressive communication in ubiquitous environments.

The aim of this article is to examine the addressing models and investigate cross-layer interactions of different mobility protocols. One of the interesting questions is how mobility should be handled and coordinated when there are multiple layers offering support for mobility. We also consider the case of the hop-by-hop routed layer-7 environment, implemented typically using SOAP (W3C, 2003), CORBA, or SIP in the telecommunications sector. These three technologies are the most frequently used, have differing characteristics and product bases, and contain the essence of middleware/application layer communication. SOAP is an abstract and generic messaging framework with extendable header system, allowing rich facilities for hop-by-hop propagation of messages.

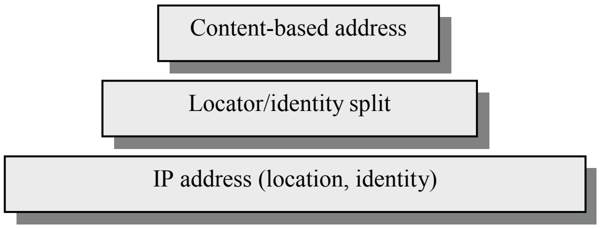
ADDRESSING MODELS

The way mobile and stationary nodes are addressed is crucial in how mobility is supported in a distributed system. We define three different addressing models for mobile systems (see Figure 1):

1. **Address with Both Location and Identity:** This form of addressing couples the communicating end-points to specific locations in a network. For example the IP address is used in both identifying a node and routing packets to it. This form of addressing typically uses a mediating stationary node to handle the mobility management and location updates for the mobile nodes.
2, **Address with Locator/Identity Split:** This way of addressing separates the identity of a node and the location of the node. This allows more flexible mobility support since the identity may be used to lookup the physical location of a node. For example the Internet Indirection Architecture (i3) and the HIP are based on this form of addressing.
3. **Content-Based Addressing:** This goes beyond locator/identity split, because it decouples the destination

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Three addressing models



from both identity and location. The destination is no longer defined by a single identity, such as the IP address or a cryptographic public key, but rather it is defined by logical rules set by applications running on the destination host. The rules are applied on messages or packets in order to make forwarding decisions. This means that using content-based addressing, we have decoupled many-to-many communication. On the other hand, the realization of content-based communication is more complex and costly. The cost of mobility in content-based routing is high when compared with the other forms of addressing. Research systems such as Siena (Carzaniga, Rosenblum, & Wolf, 2000) and Rebeca (Mühl, Ulbrich, Herrmann, & Weis, 2004) use content-based addressing.

These addressing models are not orthogonal and may be applied on different layers of the communications stack. Since the current Internet is based on the IP protocol, it provides the baseline addressing with location and identity contained in the IP address. Above that, we may implement the locator/identity split using HIP or an overlay network such as i3. Content-based addressing is also implemented above IP using application-level routers.

Identity-based mechanisms may be extended to support anonymous communication and multicast. For example, i3 supports multicast using triggers and anonymity by chaining private and public triggers. Content-based routing may, on the other hand, be extended to support identity-based communication by subscribing public keys, for example.

The addressing models have differing notions of the addressing space, in which addresses are defined. These differences can be used to characterize the difference between identity-based addressing and content-based addressing. The identity vector (public key) is a point in the flat one-dimensional addressing space of an overlay system. The content-based address, which is defined using a logical rule, is a subspace of a multi-dimensional addressing space. This illustrates the main difference, which is the expressiveness of the communication. In essence, for IP mobility there is a single, fixed indirection point; for locator/identity split there is a single indirection point; and with content-based there are multiple indirection points.

MOBILITY-ENABLING PROTOCOLS

A Taxonomy

Host mobility happens when a host relocates to a new location in the network, thereby possibly causing a change of the underlying IP address. Since IP addressing is tied to the location, this may cause a fundamental change in routing for the relocated host. This host relocation is commonly referred to as a *handover*. Handovers are usually divided into two main categories: *horizontal handovers* and *vertical handovers*. A horizontal handover is commonly understood as a handover that takes place within the same access network technology. A vertical handover is handover that takes place across different access network technologies (and from the host's point of view, between different networking interfaces).

There are also two ways of doing the handover: *break-before-make* or *make-before-break*. The difference of these two approaches is whether the mobility-enabling protocol or the terminal implementation (in hardware, point of view) allows creating connectivity to the new access network or router before leaving the old access network or router.

The host may also have several active IP addresses, which is called *multi-addressing*. Multi-addressing may also be used to realize *multi-homing,* which generally means that the client is connected to two independent networks for increased reliability. Multi-homing is also needed when several different access network technologies are used simultaneously. Server-side resiliency is commonly realized by connecting services to multiple network providers. This is called site multi-homing.

User *mobility* happens when a user changes the host device or access host, which causes a change in the underlying physical address of the user. The device characteristics may also change, for example when the user changes from a PDA (personal digital assistant) to a laptop. An important subcategory of user mobility is *session mobility,* which allows the relocation of user sessions from one host to another. Session mobility is an important requirement for current and future mobile applications, in which instant messaging (IM), multimedia, and voice sessions, for example, are moved from one device to another.

Service or *application mobility* happens when a service relocates or resides on a mobile host that moves. Service mobility may be triggered by factors not related with a user, for example load balancing.

Network Layer Solutions

The current solutions being standardized by IETF for network-layer mobility support are the Mobile IPv6 (MIP6) and Mobile IPv4 (MIP4) protocols. MIP is a layer-3 mobility protocol for supporting clients that roam between IP networks. Upper-layer protocols and applications are unaware of possible changes in network location and thus can operate uninterrupted while the host moves.

MIP6 mobility support consists of the triangle of the *home agent* (HA), *correspondent host* (CH), and the *mobile node* (MN). MIP4 has an additional optional networking node called *foreign agent* (FA) that has been left out from the MIP6 specifications. In both MIP versions the HA serves as an anchor point for MNs, and any CN may communicate and initially reach the MNs through the HA.

The basic MIP routing is triangular. A CN sends packets to an MN via an HA, and then the HA tunnels packets to MN's current location. Finally the MN sends packets directly to the CN. In practice, triangular routing is inefficient and generally also impossible due to widely used *ingress filtering.* Practical MIP deployments either route all packets via HA (reverse tunneling) or the MN and CN communicate directly (route optimization, which can be negotiated with the help of HA). The distance between the MN and the HA may also be long both topologically and geographically. Thus routing packets between the MN and the HA may cause considerable delay. However, to improve the situation, an HA may also be allocated from the network (Calhoun, Johansson, Perkins, Hiller, & McCann, 2004) the MN is currently visiting. A similar way of optimizing IP mobility is utilizing some form of localized or hierarchical mobility management.

The *hierarchical mobile IPv6 mobility* (HMIPv6) (Soliman, Castelluccia, Malki, & Bellier, 2005) management solution introduces local *mobility anchor points* (MAPs) that are essentially Home Agents. MAPs can be located at any level in a hierarchical network of routers, including the access routers. The aim of the HMIPv6 is to minimize the signaling latency and reduce the number of required signaling messages. As long as the MN stays inside one MAP domain, it only needs to update its location with the MAP. The localized mobility management can also be completely handled on the network side without MN's involvement at the IP mobility protocol level. In these cases the network side needs to employ some kind of tunneling or local routing solution that is transparent to the MN.

Another network-layer mobility solution being standardized by IETF is the NEMO. The technical solution of NEMO is close to that of MIP6. NEMO allows complete sub-networks to change their location in a network instead of single hosts. This is realized with a mobile router that manages the mobile network. Hosts behind the mobile router do not need to be aware of mobility in any way. All packets destined to hosts behind the mobile router get routed towards the virtual home network. Then an HA managing the virtual home network tunnels all packets to the mobile router managing the mobile network.

One practical application of IPSec-based VPNs is to extend the user's home network environment to be accessible from any location. In a tunneled mode, VPNs tunnel all packets between the mobile host and a *security gateway* (that is usually located at the edge of user's home network). Until recently IPSec VPNs have not survived the change of underlying IP addresses that are also used as the outer IP addresses of the VPN tunnel. Both IKE (Kaufman, 2004) and IPSec SAs (*security associations*) had to be rekeyed after IP addresses of either end of the tunnel changed. This practically caused all existing connections to drop. Recent developments in standardization have addressed this issue. For example, MOBIKE aims to support a way to update the IKE SA and IPsec SA endpoint addresses without rekeying the SAs. This would allow keeping the existing IKE and IPsec SAs in place even when the IP address changes.

Transport Layer Solutions

The recent need for multi-homing support for transport protocols has made it possible to provide limited mobility support at the transport level. Examples of recent such transport protocols are DCCP with multi-homing and mobility extension to the base protocol (Kohler, 2006), TraSH, and mSCTP. The latter two are based on mobile SCTP, which is defined as SCTP with the ADDIP *extension* (Stewart et al., 2004).

The basic idea of transport layer mobility is to maintain the end-to-end connectivity at the transport layer, and solve the mobility problem without additional infrastructure and functionality at the network layer. When a host's point of attachment to the network changes—that is, the underlying IP address changes—the transport-layer mobility protocol needs to refresh the association between the MN and the CN using some protocol-dependent mechanism. This approach is very appealing because it requires no additional tunneling and does not interfere the natural routing of IP packets. It has also been shown that transport-layer protocols are capable of smooth handovers (Fu et al., 2004).

The biggest downside of current transport-layer mobility solutions is the lack of proper mobility management. As long as everything is MN initiated, the proposed solutions work. If a CN needs to locate an MN or both communicating ends are mobile (so called *double jump problem*), current transport-layer mobility solutions most probably fail to work properly. Proposals to solve the mobility management are still open research issues.

MIP suffers from a number of limitations, such as packet loss, high handover latency, packet encapsulation overhead, and conflict with network-level security solutions (Fu et al., 2004; Schulzrinne & Wedlund, 2000). MIP requires that location management resides on the HA. TraSH decouples location management from data traffic forwarding (Fu et al., 2004), and thus a lookup server, such as DNS, may be used for location management. TraSH leverages the multi-homing capabilities of SCTP. The difference between TraSH and MIP is that TraSH sends packets directly to the MN without using the HA.

Between Network- and Transport-Layer: HIP

Above the network level, we have various requirements for mobility in the transport and application layers. Transport-level mobility support needs to cope with changing subnets and prevent, for example, socket errors during mobility. HIP is located between the network and transport layers, and provides this kind of functionality by associating each socket to a public cryptographic key instead of an IP address. The fundamental idea behind HIP is to separate the address of a network-addressable node to two parts: the identity and locator parts. The identity part uniquely identifies the host using a cryptographic namespace, and the locator part uniquely defines the location of the node. The former part is assumed to be a long-living identifier. The latter is typically the IP address of the mobile node. Additional benefits of HIP are authentication and support for denial of service (DoS) attacks through cryptographic puzzles in the initiation phase of the protocol (Moskowitz, Nikander, Jokela, & Henderson, 2006).

Application Layer

SIP mobility support is similar in nature to the HA mechanism used in MIP. SIP mobility support is based on the *home registrar,* which is a rendezvous point for information for a particular user. SIP mobility is simplest for pre-call mobility that only requires updating of the home registrar. In addition, SIP supports midcall mobility, which requires the mobile host to send an INVITE request with the new IP address to the correspondent host (Schulzrinne & Wedlund, 2000). SIP supports session mobility, in which a media session can be maintained while changing hosts. Moreover, the end point of an active session may be changed to another device.

The Wireless CORBA specification was designed to provide a minimal useful functionality for CORBA applications. The specification defines extensions and protocols for applications in which clients and servers are executed on hosts that can move. The specification introduces the *mobile IOR* (interoperable object reference), which is a relocatable object reference that identifies the *access bridge* and the terminal on which the target object resides (OMG, 2004).

An entity called the *home location agent* (HLA) keeps track of the access bridge to which the terminal is currently connected. The mobile IOR provides mobility transparency and contains either the home location agent's address or the last known access bridge of the mobile host. In the former case the HLA will provide the new address of the mobile host. In the latter case, the last known access bridge provides the current address or forwards the invocation. Each terminal is identified using a unique terminal identifier.

The access bridges may support handoff and the specification defines two different cases of handoff: the *backward* and *forward handoff.* The former is the normal case and the latter is used to re-establish connectivity after a sudden disconnection. The backward handoff (or simply handoff) may be network initiated or terminal initiated, whereas the forward handoff is always terminal initiated.

The Internet Indirection Infrastructure (i3) (Stoica et al., 2002) is an overlay network that aims to provide a more flexible communication model than the current IP addressing (Stoica et al., 2002). In i3 each packet is sent to an identifier. Packets are routed using the identifier to a single server in the distributed system. The server, an i3 node, maintains triggers that are installed by receivers that are associated with identifiers. When a matching trigger is found, the packet is forwarded to the associated receiver. An i3 identifier may be bound to a host, object, or a session, unlike the IP address, which is always bound to a specific host.

The *robust overlay architecture for mobility* (ROAM) builds on i3 and allows end-hosts to control the placement of rendezvous points (indirection points) for efficient routing and handovers (Zhuang, Lai, Stoica, Katz, & Shenker, 2002). ROAM uses trigger server caching and trigger sampling, and supports fast handovers and multicast-based handovers for make-before-break. ROAM supports legacy applications using a user-level proxy that encapsulates IP packets within i3 packets and manages trigger-related operations.

Application mobility may require special mobility protocols, for example for applications that use or participate in overlay networks. Content-based routing and publish/subscribe (pub/sub) networks (Eugster, Felber, Guerraoui, & Kermarrec, 2003; Tarkoma et al., 2003) are examples of this kind of behavior. These systems build large-scale multicast event distribution trees over point-to-point communication links. When an application providing or subscribing certain events moves, a part of the routing topology needs to be updated to reflect this change (Burcea, Jacobsen, de Lara, Muthusamy, & Petrovic, 2004).

Pub/sub systems require their own mobility protocols in order to update the event routing topology and optimize event flow. In content-based routing of information, event brokers forward notifications based on a routing configuration established by advertisement and subscription messages. The main motivation for a pub/sub mobility protocol is the avoidance of triangle routing through a designated home broker, which may be inefficient. Experimental results shows that home-broker-based approaches do not perform well (Tarkoma, Kangasharju, & Raatikainen, 2003; Burcea et al., 2004). Mobility protocols are also needed for load balancing subscribers and advertisers between brokers. Efficient mobility protocols for pub/sub are currently an active research topic.

Table 1. Differences in mobility-aware systems and protocols I

	MIP	HIP	SIP
Target	MN	MN	Session
Mechanism	HA	DNS/Overlay	Home registrar
Buffering	No	No	Yes (stateful proxy)
Update Point	1 Fixed	1	1 Fixed
Location Privacy	Yes (not /w route opt)	No	No
Authentication	Yes	Yes	Yes

Table 2. Differences in mobility-aware systems and protocols II

	WCORBA	i3	Pub/Sub
Target	Object	Any	Subs/Advs
Mechanism	Home bridge	Rendezvous point	Hop-by-hop
Buffering	Yes	No	Yes
Update Point	1 Fixed	1	>1
Location Privacy	Yes	Yes	Yes
Authentication	-	-	-

DISCUSSION

Tables 1 and 2 illustrate the differences between different mobility-aware systems and protocols discussed in this article. The target denotes the nature of the mobile entity. The target is the mobile node for MIP and HIP. Higher-level mobility protocols allow more fine-grained mobility, SIP supports session mobility, WCORBA object mobility, i3 leaves the entity unspecified, and mobile pub/sub systems support the mobility of subscriptions and advertisements to various degrees.

The *mechanism* term denotes the type of handover protocol, for example the HA- based scheme of MIP or DNS/Overlay update of HIP. SIP uses the home registrar for location updates, WCORBA has the home bridge. The i3 overlay uses a rendezvous point that manages the triggers, and mobile pub/sub systems typically have to update the whole routing path between the source and destination of mobility. The update point denotes the number of indirection points and whether or not they are fixed. MIP, WCORBA, and SIP have a single, fixed indirection point: the HA, home bridge, or the home registrar. HIP—with overlay-based address resolution—and i3 have a single, non-fixed indirection point. Finally, pub/sub systems have typically multiple indirection points.

Buffering of packets and messages is a useful functionality for supporting disconnected operation. MIP and HIP do not support this feature. SIP supports disconnected operation through stateful proxies and buffering messages for delivery. SIP also copes with network partitions using retransmission (Schulzrinne & Wedlund, 2000). WCORBA access bridges maintain a list of pending invocations, and pub/sub systems buffer notifications for disconnected clients. The i3 overlay, on the other hand, does not buffer packets.

Location privacy hides the current IP address of the mobile entity. MIP supports location privacy with the exception of the route-optimization option. HIP and SIP do not support location privacy. A HIP host has access to the IP address corresponding to the public key (host identity) of a mobile node. SIP does not hide the IP address; it is disclosed in the session description. WCORBA provides support for location privacy. Terminals are addressed using the terminal identifier, and the access bridge hides the transport address of the terminal. On the other hand, the location of an access bridge is revealed. The i3 overlay supports the hiding of source addresses with private triggers. Typically mobile pub/sub systems, such as Siena and Rebeca, provide anonymous communication, and only the edge brokers know the transport addresses.

Authentication of terminals and users is also an important functionality for mobile systems. MIP provides authentication of signaling messages using various authorization extensions. IPSec and Internet Key Exchange (IKE) can be used to protect the integrity and authenticity of MIP signaling. On the other hand, IPSec and IKE were not initially designed for multi-homed operation, and currently, multi-homed operation has overhead due to additional database entries and key negotiations for each pair of source/destination address. HIP supports authentication through the host identity, which is essentially a public key. SIP supports three authentication styles: HTTP-style basic authentication, digest authentication, and S/MIME. WCORBA may be used with the CORBA security specification (OMG, 2002), and i3 and pub/sub systems may be extended to support various forms of authentication.

CROSS-LAYER INTERACTIONS

Mobility support is currently available on many layers. Assuming that the base network-level routing technology is the Internet protocol (IP), mobility solutions are also needed on many layers. First, the lower layers need to be able to detect mobility and activate higher-level protocols. These mechanisms are outside the scope of this article. Second, after physical mobility the IP address will change, and after user mobility it may change; both instances require solutions for informing others that the address of the host has changed. This may be accomplished by pure network-level solutions (mobile IP), a hybrid approach (HIP), or through purely application-level solutions (overlay, DNS update). The mobility of sessions and objects is more subtle than the mobility of hosts, and we need mechanisms above L3 to cope with session and object mobility due to reconfiguration and, for example, load balancing decisions.

Lower-level mobility protocols are in some cases complementary to higher-level protocols. Both MIP and HIP are complementary, as is the case for MIP and middleware mobility protocols—for example, WCORBA and multi-hop application-level overlays such as i3 and Siena. If SIP or WCORBA provides mobility support, a higher-level mobility solution is not required. In general, a higher-level handover protocol is required for more fine-grained mobility support and optimizations. If only host mobility support is required, higher-level protocols are not needed. MIP may still be useful in scenarios where application-level mobility protocols are used and the client needs to be contacted using the home IP address. For example, pub/sub and many overlay systems hide the IP addresses of the communicating entities. In these systems the IP address of the terminal (care-off address) is not used for end-to-end communication, but only for the last-hop at the edge of the network. If MIP is not used, clients cannot address the mobile node directly since its IP address changes with the access location.

If MIP or HIP are not used, higher-level protocols need to provide mobility support. In this scenario there are no interactions between network-level mobility protocols and higher-level protocols. The transport, middleware, or overlay layer handles the locator/identity split and indirection. There are many possibilities for supporting mobile applications, for example SCTP, mSCTP, TraSH, and the middleware protocols.

HIP provides useful features for the higher layers, such as persistent transport-layer connections, multi-homing, and authentication. On the other hand, the timeliness and efficiency of the HIP key/address distribution mechanism is still open and requires interaction with an application-level lookup service, such as DNS or i3. The use of HIP with SCTP, SIP, and other protocols are currently open issues.

Each networking layer operates mostly on its own. Due the lack of cross-layer interactions, similar tasks may be repeated multiple times on each layer. We can consider access authentication and service authorization as an example of this. Several networking layers, including the final application-level service, may initiate similar authentication and authorization transactions that all end up at the same AAA (*authentication, authorization, accounting*) backend system located in the user's home network or the service provider's premises.

CONCLUSION

We presented taxonomy of mobility-enabling protocols on different layers of the networking stack starting from the network layer. Three useful addressing models were identified that are inherently related with mobility: addresses with both location and identity, locator/identity split, and content-based addressing. The first is currently used in the IP network architecture, and mobility solutions typically use triangle routing with optimizations. The second model has been proposed, because of its mobility-friendly characteristics—mainly a non-fixed point of indirection. The last model has also been proposed for expressive communication and is believed to be a good candidate for supporting mobile and distributed applications.

The distinguishing feature of the models is the number of indirection points and whether or not they are fixed. Content-based addressing is most flexible, but may also require the update of several indirection points, which is more costly than updating a single point. We identified several useful cross-layer interactions, but noted that some protocols are complementary. Especially if middleware mobility support is used, the benefit of using lower-level mobility protocols is uncertain. Middleware mobility support is needed for object, session, and pub/sub mobility.

REFERENCES

Burcea, I., Jacobsen, H.-A., de Lara, E., Muthusamy, V., & Petrovic, M. (2004). Disconnected operation in publish/subscribe middleware. *Proceedings of the IEEE International Conference on Mobile Data Management.*

Calhoun, P. R., Johansson, T., Perkins, C. E., Hiller, T., & McCann, P. J. (2004, August). *Diameter Mobile IPv4 application.* Standards Track RFC 4004.

Carzaniga, A., Rosenblum, D. S., & Wolf, A. L. (2000, July). Achieving expressiveness and scalability in an Internet-scale event notification service. *Proceedings of the 19th ACM Symposium on Principles of Distributed Computing (PODC2000)*.

Devarapalli, V. et al. (2004, June). *Network mobility (NEMO) basic support protocol.* Standards Track RFC 3963.

Eugster, P.T., Felber, P. A., Guerraoui, R., & Kermarrec, A.-M. (2003). The many faces of publish/subscribe. *ACM Computing Surveys,* (2), 114-131.

Fu, S., Atiquzzaman, M., Ma, L., Ivancic, W., Lee, Y.-J., Jones, J. S. et al. (2004, January). *TraSH: A transport layer seamless handover for mobile networks.*

Johnson, D., Perkins, C., & Arkko, J. (2004, June). *Mobility support in IPv6.* Standards Track RFC 3775.

Kaufman, C. (2004, September). *Internet Key Exchange (IKEv2) protocol.* Standards Track RFC 4306.

Kivinen, T. (2006, March). *Design of the MOBIKE protocol.* Internet Draft.

Kohler, E. (2006, January). *Datagram congestion control protocol mobility and multi-homing.* Internet Draft.

Mühl, G., Ulbrich, A., Herrmann, K., & Weis, T. (2004). Disseminating information to mobile clients using publish/subscribe. *IEEE Internet Computing,* (May), 46-53.

Moskowitz, R., Nikander, P., Jokela, P., & Henderson, T. (2006, March). *Host identity protocol.* Internet Draft.

OMG. (2002). *CORBA Security Service v.1.8.* Object Management Group.

OMG. (2004, April). *Wireless access and terminal mobility in CORBA v.1.1.* Object Management Group.

Perkins, C. (2002, August). *IP mobility support for IPv4.* Standards Track RFC 3344.

Schulzrinne, H., & Wedlund, E. (2000). Application-layer mobility using SIP. *SIGMOBILE Mobile Computing Communication Review*, 4(3), 47-57.

Soliman, H., Castelluccia, C., Malki, K. E., & Bellier, L. (2005, August). *Hierarchical Mobile IPv6 mobility management (HMIPv6).* Experimental RFC 4140.

Stewart, R. et al. (2004, June). *SCTP dynamic address reconfiguration.* Internet Draft.

Stoica, I., Adkins, D., Zhuang, S., Shenker, S., & Surana, S. (2002). Internet indirection infrastructure. *Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (pp. 73-86). ACM Press.

Tarkoma, S., Kangasharju, J., & Raatikainen, K. (2003). Client mobility in rendezvous-notify. *Proceedings of the International Workshop on Distributed Event-Based Systems (DEBS'03)*.

W3C. (2003, June). *SOAP version 1.2.* W3C Recommendation.

Xing, W., Karl, H., Wolisz, A., & Müller, H. (2002, October). M-SCTP: Design and prototypical implementation of an end-to-end mobility concept. *Proceedings of the 5th International Workshop on the Internet Challenge: Technology and Applications,* Berlin, Germany.

Zhuang, S., Lai, K., Stoica, I., Katz, R., & Shenker, S. (2002). *Host mobility using an Internet indirection infrastructure.* Technical Report, University of California at Berkeley, USA.

KEY TERMS

Break-Before-Make: A mobility solution feature where the connection to the old point of attachment must be torn down before establishing a connection to the new point of attachment to the network during handover.

Content-Based Routing: The process of forwarding messages based on their content.

Horizontal Handover: A handover within the same access technology.

Locator/Identity Split: Form of addressing, in which the identity and location of a node have been separated.

Make-Before-Break: A mobility solution feature that allows establishing connectivity to the new point of attachment in the network prior to tearing down the connection to the old point of attachment during handover.

Mobile IP: A classical layer-3 mobility solution based on tunneling and a stable anchor point representing the mobile node.

Service or Application Mobility: The mobility of an application or service from one physical location to another.

Session Mobility: The seamless transfer of an ongoing communication session from one device to another.

User Mobility: The ability of an end user to send and receive information regardless of mobile terminal and current location.

Vertical Handover: A handover between different access technologies.

Using Mobile Devices for Electronic Commerce

Raul Fernandes Herbster
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

Electronic commerce is becoming the most used mechanism for non-traditional commerce. However, several popular delivery services are still accessed via telephone, which enables commerce anytime, anywhere. Such telephony-based services have several problems: they do not offer a more detailed description of available products; users may ask the attendant to repeat the description of a certain product, directly affecting the time of product selling; the number of concurrently attended clients is limited to the number of attendants; and the product list must also be continuously updated, by adding or removing products, but the user cannot be automatically informed about that.

Mobile devices offer a sophisticated interface that allows better user interaction by means of lists, menus, multimedia features such as images, and much more. A user can indefinitely explore product categories very fast. It is possible to offer a more detailed description of products, with visual elements such as pictures or even videos. Besides, the number of concurrent accesses depends only on the number of connections supported by the server.

In this article, we describe an architecture for mobile commerce which allows the use of mobile devices for electronic commerce. The architecture enables the development of applications to be executed on a mobile device, which lists selling products having their own textual descriptions and pictures. We discuss architectural modules and the implementation of an application for selling fast food called Mobile Menu. We begin with the main background concepts related to our proposed architecture.

BACKGROUND

Electronic commerce has attracted significant attention in the last few years (Varsghney & Vetter, 2002). The continuously increasing number of users of mobile devices, such as mobile phones and personal digital assistants (PDAs), and advances in wireless network technology provide an ideal scenario for offering personalized services to mobile users and give place to the rapid development of mobile electronic commerce (MEC) (Tsalgatidou, Veijalainen, & Pitoura, 2000).

The way MEC operates is partially different from Internet e-commerce due to special characteristics and constrains of mobile terminals and wireless networks. The context, situation, and circumstances under which people use their mobile devices are also different (Tsalgatidou et al., 2000).

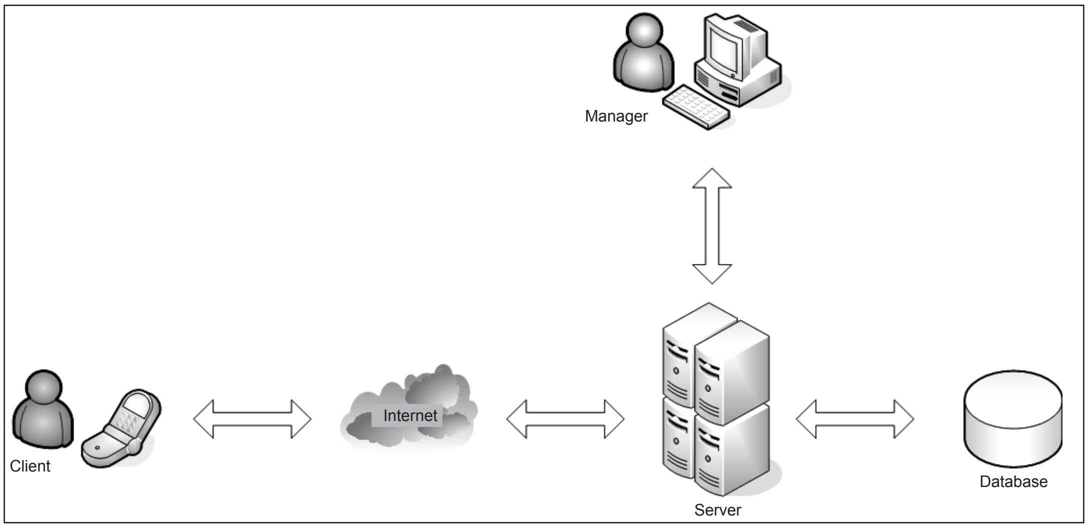
Wireless and mobile networks are increasing in exponential rate in terms of capabilities of mobile devices and user acceptance (Varsghney & Vetter, 2002). Today, more than 1 billion cell phones and other mobile devices are in use worldwide. MEC also has more advantages than traditional e-commerce applications: location-awareness, adaptivity, ubiquity, personalization, and broadcasting (Tsalgatidou et al., 2000). Applications for mobile devices are also easier to use, because the user interface of such devices is very intuitive.

Mobile devices have less resources than desktop and mainframes computers: limited memory, disk capacity, and computational power. The user interface of such devices also has some constrains: for example, small screens and small multi-functional keypads (Tsalgatidou et al., 2000). These constrains restrict the variety of applications for mobile devices and must be taken into account when designing new systems for such platforms.

Applications that demand a considerable quantity of system resources are harder to develop for mobile devices. For example, applications that need a large database to constantly perform queries and update the data are very difficult to develop for mobile devices, because the limited memory of devices does not support a database management system (DMS). Distributed architecture shares tasks among the elements of it, so that harder activities which demand memory and computational power can be allocated to those which have more resources.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Mobile menu architecture



Client-server architecture is largely used as a distributed design; it shares the tasks of the elements and provides a certain level of decoupling. It has two elements that establish communication with each other: the front-end or client, and the back-end or server. The client makes a service request to the server whereas the server provides service to the request. The client-server architecture allows an efficient way to interconnect programs that are distributed at different places (Jorwekar, 2005). However, client-server architecture is more than just a separation of a user from a server computer (Fastie, 1999). Each portion also has its own modules: presentation, which handles inputs from devices and outputs to screen display, application, and data; application, which has the rules of the business; and data, which provides services for storing the data of the application (Fastie, 1999).

AN ARCHITECTURE FOR MOBILE COMMERCE

We propose an architecture that enables mobile commerce for mobile devices. The architecture is illustrated in Figure 1. The application has three elements: the *client,* which requests the information about selling products; the *manager,* which updates information on the products; and the *server,* which receives requests and sends responses.

To start with, the user accesses the service anytime and establishes a connection with the server. Then, a list with pre-defined categories of products is sent to the user. These categories help the user to browse through the list of products. After selecting the product, the user can obtain more specific information about it or purchase the item, if more detailed information about the product is requested. At the other side, the server receives the request of purchasing or obtaining more information about the product, such as name, description, and price. Other more elaborate elements that describe the products, such as pictures and videos, can be attached. The application can be accessed anytime.

The client application can run on a mobile device and establishes a connection with the server. It requests services to the server and receives the data. The server has two modules: the network layer, which manages the network connection of the mobile devices; and the database layer, which establishes connection with the database and manages data.

Another important element of the application is the manager side. It is a desktop application which interacts with the system by modifying the database: it inserts or removes products and also modifies information about them, like price, name, and description.

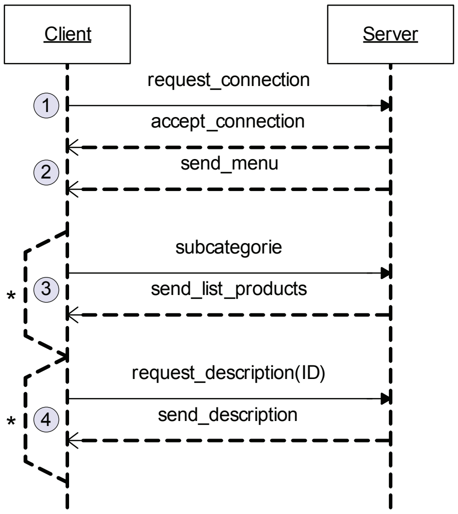
This architecture provides an interesting solution by delegating tasks for members of the system: client, server, and manager. The tasks performed by each one does not demand a considerable amount of resources from each system. For example, a mobile device cannot store a large amount of data. Thus, the architecture delegates the storing/processing tasks to the server that is supposed to have more resources. The layered architecture decouples the modules, so each one can be modified interchangeably, and also provides reuse of code.

Mobile Menu

Food delivery is a frequently used service usually accessed via telephone. It has some problems that affect the business, for example: detailed descriptions about the products are not easy available; for each client, an attendant is allocated; and a request can be mistakenly made.

Based on the architecture described, we developed Mobile Menu, which is an application for fast food mobile commerce.

Figure 2. Client-server protocol



It was implemented using Java™ Platform–J2SE for desktop applications and J2ME for mobile applications.

The functioning of Mobile Menu is represented in Figure 2. The client connects to the server and requests the menu, which is sent by the server (1, 2). The first part of the menu has only categories and subcategories of the products (pastas, salads, drinks, beverage, etc.) (3). If the client wants to see the products of a specific subcategory, it sends a request to the server and the latter sends a list of all products (3). This list contains the name of each product and its ID. If the user wants to know a detailed description of the product, the client sends a request to the server and receives additional information (4).

The menu can be updated anytime through the manager side of the application. Mobile Menu adds considerable business value in food delivery services, usually accessed via telephone. The user can indefinitely explore the categories very quickly. It is possible to offer a more detailed description of products, with visual elements like pictures or even videos on them. The number of accesses depends only on the quantity of connections supported by the server.

FUTURE TRENDS

Mobile e-commerce has several classes of applications (Timmers, 1999) with different characteristics. For example, Internet banking is a very different kind of mobile commerce which is not explored here. As for future works, we propose enhancing the architecture so more classes of mobile applications can use it.

For example, the architecture does not solve problems like selling multimedia products such as video to mobile devices. These devices have limited memory/processing. Thus, storing and playing large videos with high quality, for example, is not viable. Another aspect that was not explored is data security: the messages are not encrypted and the channel is not secure. The architecture can be adapted by adding other modules to solve some security problems.

CONCLUSION

Mobile e-commerce is an area that creates opportunities for many players in the field. Mobile devices have some constraints that must be taken into account when designing the application for such platforms. However, these devices also have important characteristics that make them an interesting channel of commerce.

The number of services available through mobile devices is increasing: banking, purchasing of images and music, and much more. These applications make services easily accessible to the client, providing consumption of a large variety of products and services in a convenient way.

In this article we proposed an architecture for mobile devices that enables the use of mobile devices for electronic commerce. We applied the proposed architecture to develop Mobile Menu, an application for food delivery based on mobile commerce. Although this application is specific for such a domain, the architecture is generic and could be adapted to any mobile commerce application.

REFERENCES

Fastie, W. (1999). Understanding client/server computing. *PC Magazine,* 229-230.

Jorwekar, S. (2005). Client server software architecture.

Pitoura, E., & Samaras, G. (1998). *Data management for mobile computing.* Kluwer Academic.

Timmers, P. (1999). *Electronic commerce: Strategies and models for B2B trading.* New York: John Wiley & Sons.

Tsalgatidou, A., Veijalainen, J., & Pitoura, E. (2000). Challenges in mobile electronic commerce. *Proceedings of the 3rd International Conference on Innovation Through E-Commerce,* Manchester, UK.

Varsghney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7,* 185-198.

Varsghney, U., & Vetter, R. (2000). Emerging wireless and mobile networks. *Communications of the ACM.*

Wesel, E. K. (19998). *Wireless multimedia communications, networking video, voice and data.* San Francisco: Addison-Wesley.

KEY TERMS

Client-Server Architecture: A basic concept used in computer networking, wherein servers retrieve information requested by clients and clients display that information to the user.

Distributed Network: A system where resources are spread among many computers, instead of being stored in a single location

Electronic Commerce (E-Commerce): The buying and selling of information, products, and services electronically over the Internet.

Mobile Devices: Any portable device used to access a network (Internet, for example).

Multimedia Application: Applications that support the interactive use of text, audio, still images, video, and graphics.

Protocol: A set of rules and procedures governing communication between entities connected by the network.

User Interface: The means by which an individual communicates with a computer through a software application. The common methods for such communication are, commands, menus, and icons.

Wireless Network: Networks without connecting cables, that rely on radio waves for transmission of data.

Using Mobile Devices to Manage Traffic Infractions

Stefânia Marques
Federal University of Campina Grande, Brazil

Sabrina Souto
Federal University of Campina Grande, Brazil

Miguel Queiroga
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

Mobile computing is one of the recent technologies with the most impact on people's lives. Several research and industrial applications are benefiting from mobile computing, supporting various human daily activities. Transit law enforcement officials can benefit from the availability of powerful mobile devices, such as smart phones and PDAs, to help them to execute their daily tasks. In such a scenario, an official can verify a driver's data record and issue tickets online.

In this article we describe the SM-FIT system that makes it possible for transit law enforcement officials to perform online queries about potential infractions of a driver of a vehicle by using a mobile device. Queries are performed based on a unique identifier: the driver's license number.

The system is implemented based on a client-server paradigm, where mobile devices are clients and servers are base stations. Clients must have a local database to store each result of a query, when needed. Each registry stored has the following attributes: a unique identifier, the number of the vehicle's plate, the date and time that the officer registered the infraction, and the status of the infraction. Besides, photographs can be stored, digitally signed, and transmitted to a database for future prosecution.

The remainder of this article is organized as follows. We first outline some background concepts related to the system's development. We then present the proposed system architecture and functioning, and discuss some trends related to future research in this area. We close with some final remarks.

BACKGROUND

This section describes briefly some concepts related to J2ME and, more specifically, MIDlets. Such technologies have been used for developing the proposed system.

J2ME

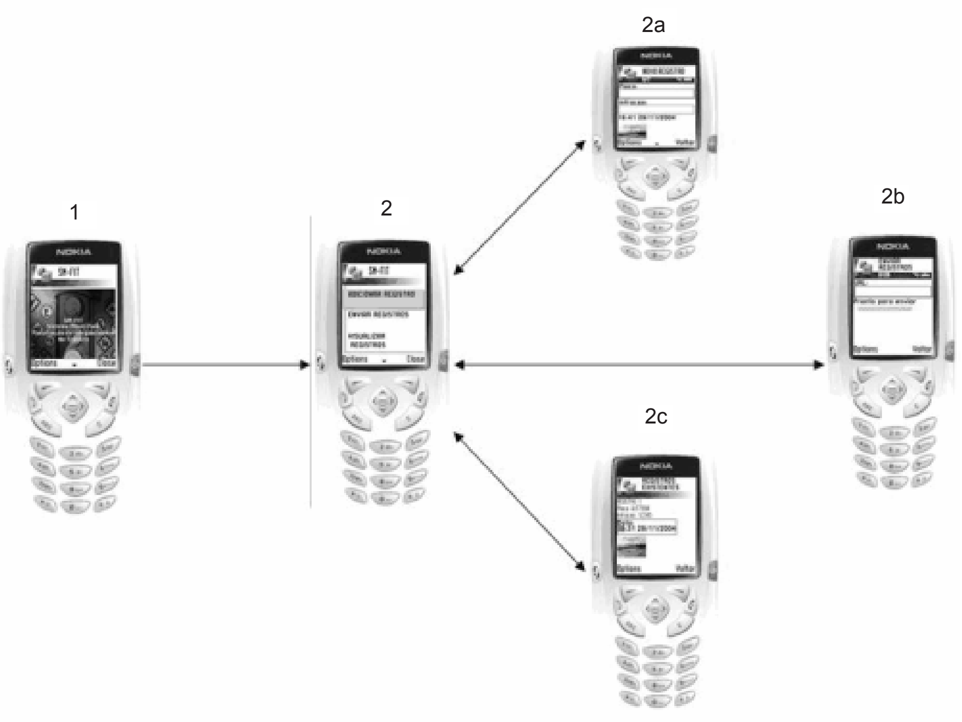
J2ME is a development platform based on Java Technology for developing mobile and embedded applications. It focuses on two types of devices:

- **High-End Consumer Devices:**
    - CDC (connected device configuration);
    - interactive TVs, videophones, wireless devices;
    - a large variety of user interfaces;
    - typical memory of 2 to 4 Mb; and
    - persistent connection, generally TCP/IP.
- **Low-End Consumer Devices:**
    - CLDC (connected limited device configuration);
    - cell phones, bidirectional pagers, PDAs, and so forth;
    - limited processors (8 to 32 MHz);
    - limited memory;
    - lazy connection, intermittent (9600bps) and generally not based on TCP/IP; and
    - powered by batteries.

The J2ME platform includes flexible user interfaces, a robust security model, a broad range of built-in network

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1.



protocols, and extensive support for networked and off-line applications. Besides this, applications based on J2ME specifications are written once for a wide range of devices.

MIDlets

Java applications running on MIDP devices are known as MIDlets, which consist of at least one Java class and have to be derived from the abstract class javax.microedition.midlet.*MIDlet*. These MIDlets use an execution environment within the Java Virtual Machine to control the application's lifecycle through a set of methods implemented by this MIDlet.

MIDlets can also use methods to obtain services from the environment. A group of related MIDlets can be put together in a MIDlet suite, which is packaged and installed in (or removed from) a device as a unique entity. MIDlets in a suite share all static and dynamic resources in their environment:

- Execution data can be shared by MIDlets, and the usual Java conventions of synchronization can be used to control data access.
- Persistent data can also be accessed by all MIDlets in a suite.

All files in a MIDlet suite must be within a JAR package. These packages contain the classes of the MIDlet and other resources, like images, and a manifest file. This manifest file contains a list of attributes and definitions to be used by application managers to install the JAR files in the device.

Security in MIDlets

The JAVA security model in its standard edition (J2SE) is too expensive in terms of costs for memory allocation, and it requires configuration knowledge that is not present in users of mobile devices. Thus, neither CLDC nor MIDP include these functionalities.

Cryptography of public key and certifiers are not available as default, so it is necessary to pay attention when installing MIDlets and, preferentially, only accept software from trustable fonts. MIDP 2.0 included the https protocol that helps to diminish these problems.

SYSTEM DESCRIPTION

The SM-FIT is a system that makes it possible for transit law enforcement officials to register transit irregularities in a local database. Each record of this database consists of the number of the vehicle's plate, a code of the infraction, date and time that the infraction occurred, and the photography of the vehicle involved in the corresponding infraction.

The system is implemented based on a client-server paradigm, where mobile devices are clients and servers are base stations. At the end of the day, transit law enforcement officials send the registers of their mobile devices to the server in order to make available more space in their devices for local databases.

Consider Figure 1 to understand the system functioning. Through the starting screen of the system (1), it is possible to access the system main menu (2). This menu of options includes: *Add a record, Send records,* and *View records.*

If the user chooses the first option (2a), to add a record, it is shown a screen with the following data to be filled: the vehicle's plate involved in the infraction, the code of the infraction, date and time of the infraction (this information is taken automatically from the mobile device), and the photography of the vehicle involved in the infraction.

If the user wants to send all the records stored in the mobile device to the server (2b), it is shown a screen that requests the IP address of the server. Then, the process of sending the information to the database is initiated.

Finally, if the user chooses to visualize the records of the mobile device (2c), it is shown a screen with a list of records contained in the local database. This visualization is taken in a way that each record is shown individually on the screen.

FUTURE TRENDS

Future research and development of new technologies may make possible the transferring of streams from mobile devices to the server. In this way, the system described here could store videos of infractions or even the voice of the transit law enforcement official describing how the infraction occurred. For this kind of improvement, it would be necessary to consider efficient lower battery consumption mechanisms.

Another point to be considered is cryptography in mobile devices, which would increase security during the transference of records from the mobile devices to the server. Considering that it is an industrial application, security is an essential requirement.

In the context of the driver, a module for drivers to consult the situation of their licenses or cars could be made available. This module would make available all the information contained in the server database.

CONCLUSION

The usage of mobile devices to manage traffic infractions can bring some benefits, such as saving time on queries made to check if drivers have any previous infractions. Another benefit is that using a mobile device that can take pictures, when the transit law enforcement official registers a new infraction, he/she can also take a picture of the vehicle involved in the infraction to prove in the future that it really occurred.

Regardless of application domain, there is a widespread use of mobile devices and an increasing need for real-time answers wherever a person is. This makes it necessary to use a technology like the SM-FIT system. It allows for transit law enforcement officials to get real-time answers for their queries about the drivers' situations, and also about their vehicles. But this system can also be adapted to be used in many other application scenarios, such as to make restaurant reservations, access credit card accounts, and make payments.

REFERENCES

Alves, D. (2004). *Introdução ao J2ME.* Retrieved November 30, 2004, from http://www.conexaojava.com.br/conexaojava04/download/minicursos/Java2.Micro.Edition-Conexao.Java.2004.pdf

Borges, R. L. (2004). *J2ME na prática.* Retrieved November 30, 2004, from http://www.ucb.br/java/JavaDays/J2ME_RosfranBorges.pdf

Easy Process. (n.d.). *A software development process.* Retrieved from http://dsc.ufcg.edu.br/~yp

Gomes, H. M. (2004). *Arcabouços de software para desenvolvimento de aplicações embarcadas.* Retrieved November 30, 2004, from http://www.dsc.ufcg.edu.br/~hmg/disciplinas/nokia/ASDAE.pdf

XP1. (n.d.). *A software development process.* Retrieved November 30, 2004, from http://www.dsc.ufcg.edu.br/~jacques/cursos/2002.2/projii/xp1/xp1.html

KEY TERMS

Base Station: A centralized repository for the storage and management of information, organized for a particular area.

CDC: Connected device configuration.

CLDC: Connected limited device configuration.

IP Address: An Internet protocol address attributed to a client or a server in the client-server paradigm.

J2ME: Java Second Micro Edition.

MIDP: Mobile information device profile.

Personal Digital Assistant (PDA): A handheld device that combines computing, telephone, Internet, and networking features.

Using Service Proxies for Content Provisioning

Panagiotis Kalliaras
National Technical University of Athens, Greece

Anthanasios-Dimitrios Sotiriou
National Technical University of Athens, Greece

INTRODUCTION

In modern broadband mesh networks, communication between two end nodes is carried out not directly, but through a number of intermediate nodes. While these nodes' only function may be to relay information from one point to another, they may also host computational elements which perform some service on behalf of other applications. We deal with the problem of optimally mapping multimedia content transcoding service elements onto network resources. There may be several places in the network where the required compression and decompression services could be performed. We would like to select the best locations that meet the application's requirements. We propose a new approximation algorithm for constrained path optimization, which provides better scalability and simplicity than previous approaches. This is accomplished basically by partitioning the overall problem into smaller ones.

RELATED WORK

The majority of the proposed schemes are focused on solving the similar multi-constrained optimal-path problem (MCOP). This problem aims to find in a network an optimal path that satisfies multiple additive path constraints and has been proven to be of NP-complete complexity, therefore unsolvable in polynomial time. Several algorithms have been proposed for the above problem. For the MCOP problem with two parameters, Jaffe (1984) proposed to use a linear weight combination of the two constraint parameters. Other proposed algorithms include Iwata et al. (1996), SAMCRA (Van Mieghem, De Neve, & Kuipers, 2001), and the Chen-Nahrstedt algorithm (Chen & Nahrstedt, 1998). All of the above algorithms require a global state to be maintained at every node. Most algorithms transform the routing problem to a shortest path problem and then solve it by Dijkstra's or the Bellman-Ford algorithm.

The concept of service path has been proposed in smaller numbers. In TranSquid (Maheshwari, Sharma, Ramamritham, & Shenoy, 2002), a transcoding and caching proxy for heterogeneous clients is proposed. In the Ninja Project (Gribble et al., 2001), service path is defined as a sequence of application-level service operators and connectors. In Lienhart, Holliman, Chen, and Yeung (2002), the authors propose the addition of a media support module (MAPS) on top of an existing peer-to-peer service layer, in order to improve multimedia services across heterogeneous computing platforms. This module is responsible for transcoding and route path selection based on the single-pair shortest-path problem and utilizes Dijkstra's algorithm to provide a solution. Our system is a combination of the above two research areas. It uses the service paths concept and also fully implements the MCOP, rather than the simpler path-finding solutions.

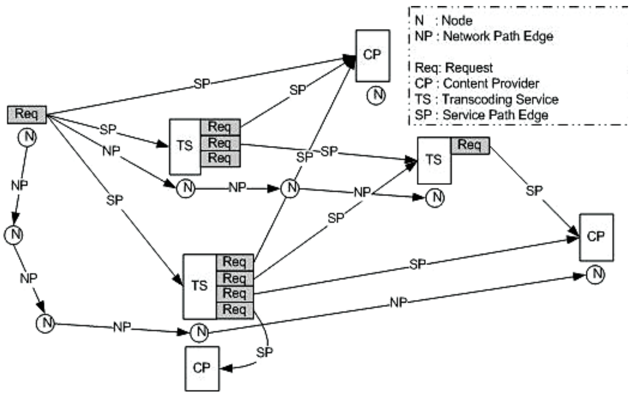
NETWORK AND SERVICE MODELING

We assume that our network topology matches that of a partial mesh network. A mesh network is reliable and offers redundancy. If one node can no longer operate, all the rest can still communicate with each other, directly or through one or more intermediate nodes. Mesh networks work well when nodes are located at scattered points that do not lie near a common line.

As a service, we denote any network resource which may include computational elements and performs some online activity on behalf of other applications. A service is an online facility that is always available to all requestors, at a predefined cost and delay. The services may be available on more than one node, either serially or concurrently. Although services in communication networks delivering multimedia content may include conversion processes like media data adaptation, merging of multiple media sources, copyright protection, metadata extraction, enhancements, and recovery, in our modeling we focus mainly on transcoding. Possible forms of transcoding include: lowering the bit rate of a media stream by reducing the image/video resolution, size, and/or frame rate; converting a media stream from one encoding format to another; or a combination of the above. The expected benefits from these adaptations are on one hand to move computation (data transformation) from the client site to the proxy, and on the other hand to reduce the volume of data transferred to the client. In any case, a service in our modeling accepts a media stream which is

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Network and service paths



characterized from an input data rate, performs application-level processing on the media data, and forwards the stream at a new output data rate.

The connection requests are initiated by clients that wish to acquire specific content from content providers. The procedure that the clients follow to detect content providers with desirable content is out of the scope of this article. The notion of clients, content providers (CP), and service providers (SP) is used here to state the dynamic nature of the network connections. The scope of multimedia service provision is to provide clients with customized and satisfactory QoS, under the constraint of end-to-end resource availability observed by each client. Several requests are initiated, targeting various service entities.

The main pattern for the proposed solution is to split the problem into smaller ones, as shown in Figure 1. Each request forms two bids. The first bid concerns the path from the request to the service that has the desired output. This bid includes the network and service paths, and the delay and cost for the path and the usage of the service. The second bid is required if the chosen service does not own the content, and acquisition is needed from somewhere else. In this case, a new request is formed and its result is returned to the initial request. The results from the second bid contain everything after the service, including any subsequent services that may be used.

PROBLEM FORMULATION AND HEURISTICS

Let directed, connected graph $G(V, E)$ denote the network topology, where V is the set of vertices of the graph (representing network components (e.g., switches, routers, hosts, aggregated subgraphs) and E the edges (representing communication links). There are four basic entities used in our problem formulation, shown in Table I. Edges and vertices are part of the actual network and services, and requests are part of the overlay network.

PROBLEM ENTITIES

Edge *e*	Vertex *v*	Service *s*	Request *req*
B_e: available bandwidth	d_v: d*elay*	v_s: host vertex	v_{req}: host vertex
d_e: delay		r_{in}/r_{out}: input/ output rate	$D_{req}^{\max}$: max d*elay*
c_e: cost per rate		c_s: cost	R_{req}: requested rate
		d_s: d*elay*	

For every request, ensure the end-to-end delay constraint is met:

$$\sum_{v \in P} d_v + \sum_{e \in P} d_e + \sum_{s \in SP} d_s \leq D_{req}^{\max} \tag{1}$$

while the available bandwidth for all edges in the path is at least the required rate at that point:

$$R_{req} \leq B_e, \forall e \in P \tag{2}$$

and minimize cost:

$$C_{tot} = \sum_{e \in P} C_e * r_e + \sum_{s \in SP} c_s \tag{3}$$

where P is the network path and SP the service path.

The problem, as described above, is more accurately described as unicast link-constrained path-constrained path-optimization routing. The link constraint refers to the available bandwidth, the path constraint to the total delay, and the desirable path optimization to the minimization of the total cost.

The Heuristics Solution

The main steps for estimating the optimal paths are the following:

- **Step 1. Topology Filtering:** For every request *req* initiated at host vertex *t* and inquiring an object with optimal rate R_{req} and maximum delay $D_{req}^{\max}$, filter all links (and possibly disconnected nodes) that do not satisfy the requested minimum linear QoS constraints, in our case, minimum available bandwidth B_e as in equation (2).
- **Step 2. Finding Available Services:** Let $S_R \in S$ be the sum of services that have an output rate R_{req}—that is,

the services that either own or can produce the object are chosen.

- **Step 3. Bid 1:** For every service $s \in S_R$ hosted at vertex t_s, solve the partial MCOP problem of (1) and (3) from t_{req} to t_s. In our implementation, we use an extended version of Jaffe's algorithm, combining: $w(P) = a * c(P) + b * d(P)$, in order to obtain the results. The main difference from the original algorithm is the usage of non-static variables *a* and *b*, which change according to the pseudocode below:

```
double α=1, b=0, step=0.1;
while (delay1 > $D_{req1}$ ‖ b!=1)      {
            α=α-step;
            b=b+step;
            P+new Dijkstra(α,b);
            delay1=getDelay(P); }
```

The logic is simple. First seek to optimize path with respect to cost only and check if the delay constraint is met. If not, increase the delay factor by a step and try again until a solution is found.

- **Step 4. Bid 2:** For every node $v \in V$ that hosts a service $s \in S_R$, if *s* has an input rate $R_{in} \neq 0$, form a new request for object with optimal rate R_{in} and maximum delay $D_{req2} < D_{req}^{\max}$.Go to step 1. Results are path2, spath2, cost2, and delay2.
- **Step 5. Negotiation of Bids:** The negotiation process deals with the sharing of partial delay constraints D_{req1} and D_{req2} between bids 1 and 2 respectively, so that delay1+delay2 < $D_{req}^{\max}$ and the overall cost is kept minimal. The negotiation is done according to the pseudocode as follows:

```
PROCEDURE bid()
            $D_{req1} = D_{req}^{\max}$ – delay2

            delay = tryBid1($D_{req1}$)

            $D_{req2} = D_{req}^{\max}$ – delay1

            delay2 = tryBid2 ($D_{req2}$)
            do
              if (delay1 < 0 && delay2 < 0)
          No feasible solution for bid1 and bid2, bid fails
              if (delay1 < 0 && delay2 > 0)
          ask trybid2 () fro a better delay, try bids again
          $D_{req2}$ = delay2 – 1
              if (delay1 > 0 && delay2 > 0)
            valid delays for both bid1, bid2, bid succeeds
               if (delay1 > 0 && delay2 < 0)
           ask trybid () for a better delay, try bids again
            $D_{req1}$ = delay1 – 1
            while bid succeeds or bid fails
END PROCEDURE
```

The worst case overall complexity of the algorithm is analogue to the complexity of the Jaffe Algorithm: $O(N\log N + 2E)$. Also, for a total number of S services, the algorithm can be run S! times, that is S times for the first time and S-k, for the k-th iteration, when s services will have been added to the service path. However, the algorithm is bounded by the maximum delay D_{req}, so given that each service s adds a mean delay $\tilde{d}_s$ (service delay + path delay), the iteration happens k=$D_{req}/\tilde{d}_s$ times and the MCOP algorithm of bid1 is executed S! / (S-k)! < S^k times. So, the overall complexity of our heuristics is O(S!/(S-k)!(Nlog N + 2E)).

SIMULATION RESULTS

The algorithm described above was implemented in Java, using the JUNG (http://jung.sourceforge.net/) software package. The simulations were conducted using three different network topologies:

- **Grid 16:** This network is composed of 16 vertices that form a grid network. Each vertex is connected to its two neighbors to the east and south, for a total of 32 edges. The east-most and south-most nodes at the edges of the square grid also have links that wrap around to the corresponding node at the opposite side, which results in a grid topology.
- **Grid 64:** This network has the same structure as the Grid 16 network described above, but has 64 vertices and a total of 128 edges.
- **Random:** The network is composed of 40 vertices and 86 edges.

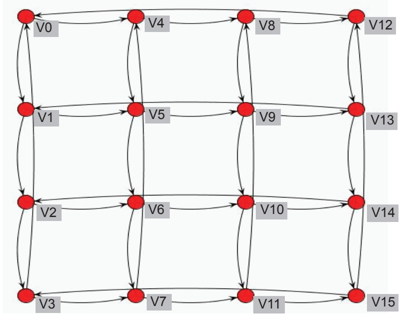
For each of the above topologies, two major sets of simulations were contacted:

- With CPs but without TSs, simulating the classic QOSR MCOP problem.
- With a variable number of service proxies. The optimal route selection heuristics described above was used in this case.

The following configuration parameters affect the simulation results:

- **Density and Location of Services:** The density of servers is defined by the ratio of the number of services to the total number of nodes. The host nodes for the services were randomly selected for Grid64 and Random, while for Grid16, all nodes hosted services.
- **Bandwidth Reserved per Request:** The bandwidth that an individual request reserves at each link is set approximately to 4% of the available link bandwidth.
- The values for the delays in edges and nodes are of the same order and 70% less than the delay caused by

Figure 2. The Grid 16 network



services, while the cost of transcoding is generally less than the cost of content provider.

Finding the Optimal Position for the Transcoding Services

The algorithm that was described previously was used in order to find the most suitable host nodes for positioning transcoding services. The Grid 16 network, shown in Figure 2, is used as an example.

The following parameters hold for nodes, edges, services, and requests:

- Every node causes a delay=1.
- Every edge has cost/rate=1, delay=4 and bandwidth=1500.
- Two services at V0 are content providers:
  - S0→10@V0 with delay=10 cost=130.
  - S0→20@V0 with delay=10 cost=20.

We wanted to find out where to put a number of TS with S10→20 and S20→10 with delay=4 cost=10 that would serve as many requests as possible. Every other node (V1-V15) was a potential host for services, so we put TSs in all of them. Next, we placed two requests for the two rates (10, 20) that were supported in our network on each node other than V0 which hosted the CPs. Their delay constraint was loose and equal to 50. In total, there were 30 requests.

As expected, from Figure 3 we see that the most useful services for implementation in our test network are the S20→10@V1 and S20→10@V4. These services compensate for the high cost of S0→10@V0. They are preferred by all 15 requests that demand rate=10, while the S0→10@V0 serves zero clients and is actually ignored.

The S20→10@V1 and S20→10@V4 services are also used for five requests that demand rate=20, despite the fact that there is already a CP (S0→20@V0) that produces it. These requests come from nodes that are far away from node V0, and it is more convenient for them to choose a service path that would reduce the cost spent at the edges of the network. For example, for node V7, the service path is SP= S0→20@V0 S20→10@V1 S10→20@V7.

From Figure 4, we confirm the fact that most network paths contain the nodes that host the most preferred services, particularly E0(V0,V1), E1(V0,V4), and E3(V1, V5).

The results from the simulations that we conducted can lead us to certain useful considerations.

Using transcoding services is necessary when there is no content provider that can satisfy a client request or its costs are unacceptable. When transcoding results to higher bit rate, the transcoder should be placed as close to the destination as possible. Accordingly, when transcoding results to lower bit rate, the transcoder should be placed as close to the content provider as possible.

Transcoding resulting in lower bit rate may be necessary even when it leads to worse performance (i.e., excess delay)

Figure 3. Importance of services

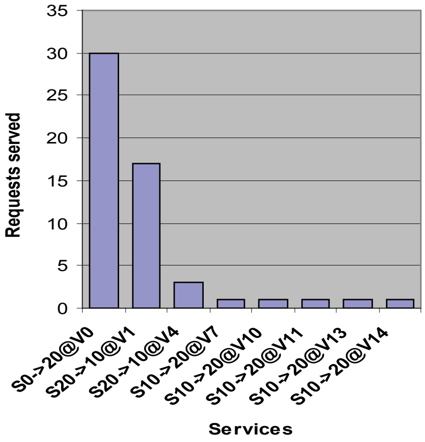


Figure 4. Importance of edges

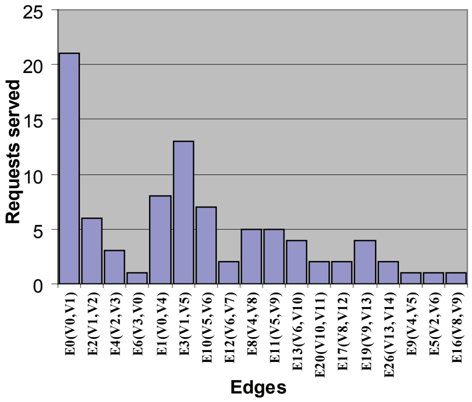
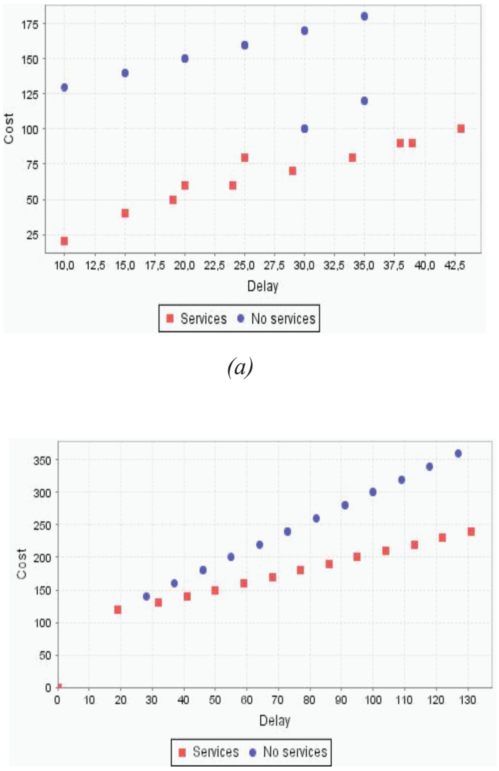


Figure 5. Cost-delay diagram for (a) Grid 16 and (b) Grid 64 networks

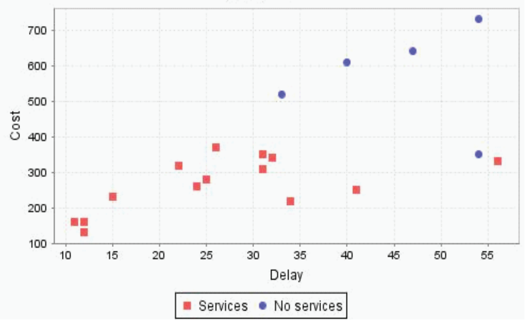


(a)

(b)

when the available bandwidth is low and many requests need to be served. Transcoding has better results when the cost and delay of the transcoding service is relatively low, and also when the host node of the service is included in the optimal path between the content provider and the receiver. In other words, the deviation of the flow for using a service must be minimal. The rate reduction is important and the request has not very tight delay constraints.

Figure 6. Cost-delay diagram for random network



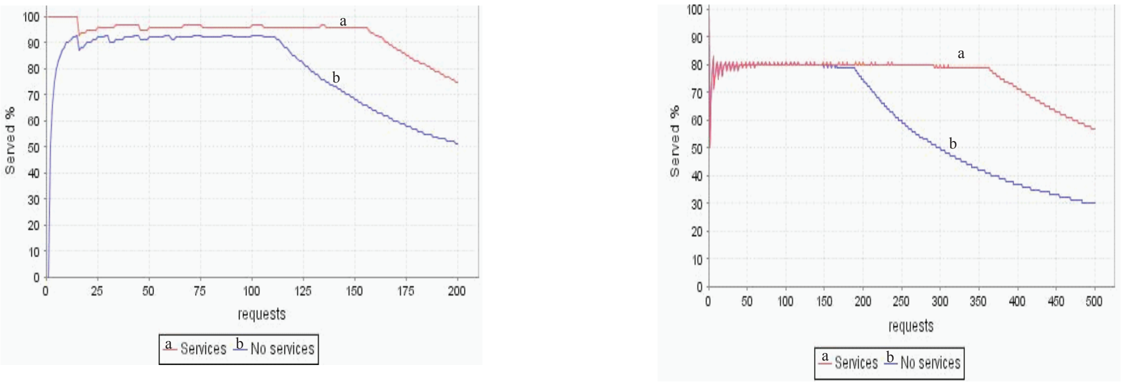
Overall Network Statistics

In the cost-delay diagrams shown in Figures 5 and 6, we show the improvements on performance by using service proxies. In all cases, only the minimum required content providers were used in the network.

In Figure 7(a) for total number of 200 requests, and in Figures 7(b) and 8 for 500 requests, the ratio of served requests is shown for each network with and without transcoding services. For the Grid 16 and Grid 64 networks, the percentages are about 90% and 80% respectively with the usage of services. Again, the benefits of using services are more evident when the network tends to reach congestion. As the optimized network resource usage increases, so does the admission control rate.

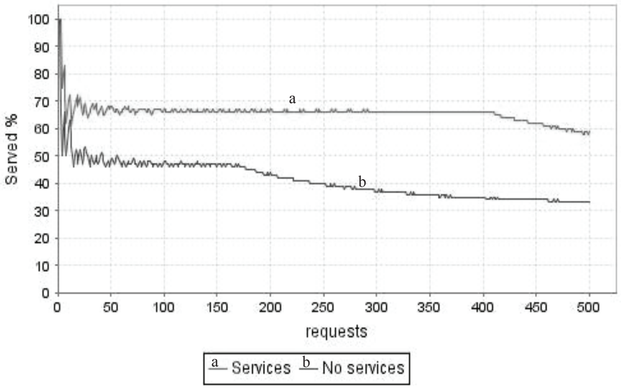
The random network has an even better comparative performance admission control rate, which is evident from the beginning due to the fact that certain client requests are unfulfilled due to the limited number of content providers.

Figure 7. Admission control diagram for (a) Grid 16 and (b) Grid 64 networks



(a) *(b)*

Figure 8. Admission control diagram for random network



CONCLUSION

In this article we have presented a distributed approach for finding the best feasible paths in networks with respect to QoS requirements. We focused on the usage of trascoding services and especially services that alter the bit rate. This approach is applied in the unicast distribution of multimedia streams. Our heuristics involved transformation of the original problem to a series of conventional shortest path QoSR problems that are solved using a new MCOP algorithm. Through our extensive simulations we proved that transcoding proxies can induce positive results in reducing costs, improving load balancing/network congestion, and consequently increasing admission control rates.

REFERENCES

Chen, S., & Nahrstedt, K. (1998). On finding multi-constrained paths. *Proceedings of ICC'98*, New York (pp. 874-879).

Gribble, S. et al. (2001). The Ninja architecture for robust Internet-scale systems and services. *Computer Networks,* (Special Issue on Pervasive Computing).

Iwata, A., Izmailov, R., Lee, D.-S., Sengupta, B., Ramamurthy, G., & Suzuki, H. (1996). ATM routing algorithms with multiple QoS requirements for multimedia Internetworking. *IEICE Transactions and Communications, E79-B*(8), 999-1006.

Jaffe, J. (1984). Algorithm for finding paths with multiple constraints. *Networks, 14*(1), 95-116.

Lienhart, R., Holliman, M., Chen, Y. K., & Yeung, M. (2002). Improving media services on P2P networks. *IEEE Internet Computing.*

Maheshwari, A., Sharma, A., Ramamritham, K., & Shenoy, P. (2002, February). TranSquid: Transcoding and caching proxy for heterogeneous ecommerce environments. *Proceedings of the 12th IEEE Workshop on Research Issues in Data Engineering (RIDE'02)*, San Jose, CA.

Van Mieghem, P., De Neve, H., & Kuipers, F. A. (2001). Hop-by-hop quality of service routing. *Computer Networks, 37*(3-4), 407-423.

KEY TERMS

Content Provider: A service that provides multimedia content.

Mobile Computing: Ability to use technology untethered, that is not physically connected, or in remote or mobile (nonstatic) environments.

Multimedia: Media that uses multiple forms of information content and information processing (e.g., text, audio, graphics, animation, video, interactivity) to inform or entertain the (user) audience.

Network: A network of telecommunications links arranged so that data may be passed from one part of the network to another over multiple links.

Network Topology: The pattern of links connecting pairs of nodes of a network.

Path Optimization: Finding a path between two vertices such that the sum of the weights of its constituent edges is minimized.

QOS Routing: Process of finding a loop-less path between nodes in a network, satisfying a given set of constraints on parameters like bandwidth, delay, etc.

Service: Any network resource that performs some online activity on behlaf of other applications.

Verifying Mobile Agent Design Patterns with RPOO

Elthon Allex da Silva Oliveira
Federal University of Alagoas - Campus Arapiraca, Brazil

Emerson Ferreira de Araújo Lima
Federal University of Campina Grande, Brazil

Jorge César Abrantes de Figueiredo
Federal University of Campina Grande, Brazil

INTRODUCTION

The act of modeling concurrent distributed systems is not a trivial task. Besides, when mobility is added to the scenario, things get worse because of new problems like variations in the communication conditions and remote execution. An important thing to be considered is how to analyze these mobile agent-based systems in order to validate and improve them.

An interesting tool to verify such systems is named Petri net. Petri net, or simply PN, is a powerful formal, graphical, and executable tool commonly used for verifying communication protocols and concurrent systems. A variant of Petri net, named RPOO (Guerrero, 2002)—Object Oriented Petri net, brings all the advantages of Petri nets and object-oriented semantics and puts them together. This union produces a new tool, with the formal semantics of Petri net models and the object-oriented semantics of some programming languages, like C++ and Java. This makes the semantics of formal behavior models closer to the semantics of those programming languages. With this, formal models of programs are built easier than with classical Petri nets. Besides, a closer model, speaking about semantics, is used to verify and analyze mobile agent design patterns which gives a more trusted verification.

This article presents the formalization and analysis of three migration design patterns—itinerary, star-shaped, and branching—done by using of RPOO. A brief comparison between RPOO models and classical Colored Petri net (Jensen, 1992, 1997) models is also briefly presented.

BACKGROUND

Mobile Agent Migration Design Patterns

In this article we consider three migration design patterns proposed in Tahara, Ohsuga, and Honiden (1999): itinerary, star-shaped, and branching patterns. In the sequence we detail each pattern. We used a message sequence diagram to show an overall picture of the design patterns.

Itinerary

This pattern provides a way to execute the migration of an agent, which will be responsible for executing a given job in remote hosts. The agent receives an itinerary on the source agency, indicating the sequence of agencies it should visit. Once in an agency, the agent executes its job locally and then continues on its itinerary. After visiting the last agency, the agent returns to its source agency. This pattern is a good solution to agents that need to execute sequential jobs. In Guedes, Machado, and Medeiros (2003) and Medcraft (2003), case studies that apply this pattern are shown.

In Figure 1, we present a possible execution sequence for this pattern. We use a notation that is equivalent to the one presented in Klein, Rausch, Sihling, and Wen (2001). In this notation, an object is used to represent an entity that controls agents' execution in a given agency (creation, destruction, migration) and indicates their location. Migrations are represented by message passing from one agency entity to the other. The message is labeled as MIGRATING AGENT. Before migration, agent execution is interrupted (arrow labeled as destroy()). Execution is continued in the target agency (arrow labeled as initialize()).

As we can see, there are three agencies: a SourceAgency and two search agencies (DestinationAgency1 and DestinationAgency2). Following the diagram, we see that there is an agent (ItineraryAgent) that sets its itinerary, moves to the first search agency where it executes its job, then it moves to the second one, executes the job, and returns to the source agency.

Star-Shaped

On the star-shaped pattern, the agent receives a list of agencies that it has to migrate to. Initially, the agent migrates to

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Message sequence chart for Itinerary

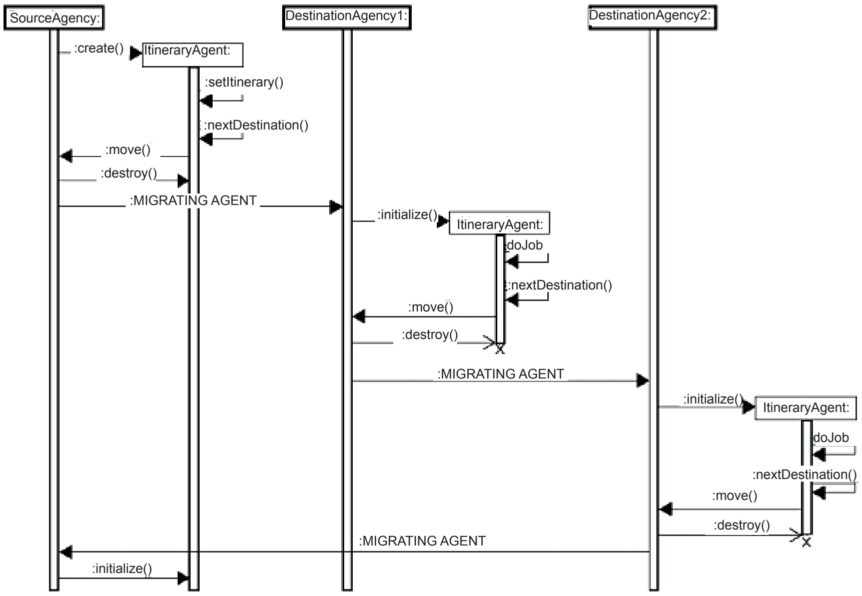
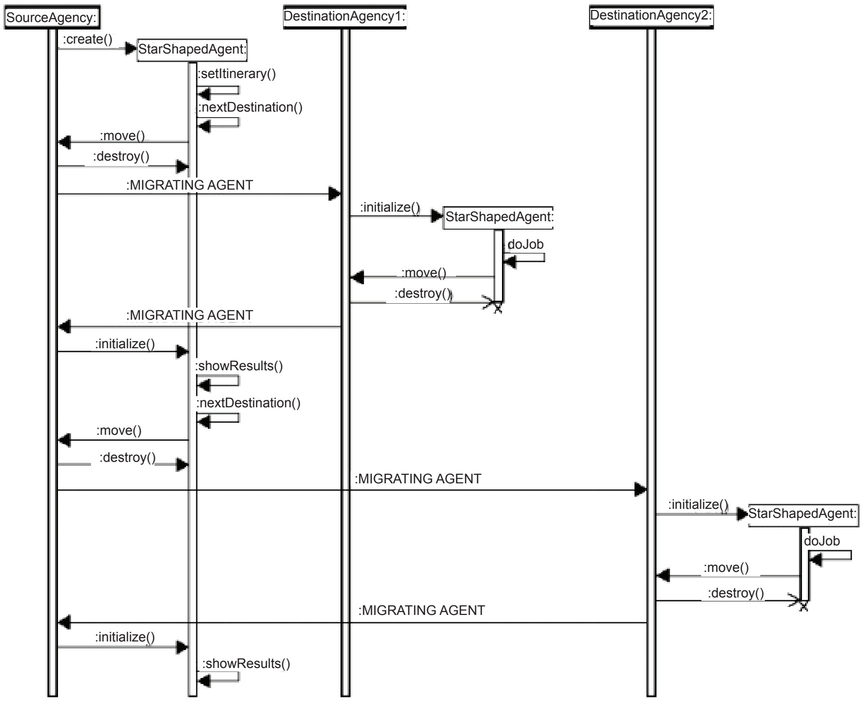

SourceAgency:
DestinationAgency1:
DestinationAgency2:
:create()
ItineraryAgent:
:setItinerary()
:nextDestination()
:move()
:destroy()
:MIGRATING AGENT
:initialize()
ItineraryAgent:
doJob
:nextDestination()
:move()
:destroy()
:MIGRATING AGENT
:initialize()
ItineraryAgent:
doJob
:nextDestination()
:move()
:destroy()
:MIGRATING AGENT
:initialize()

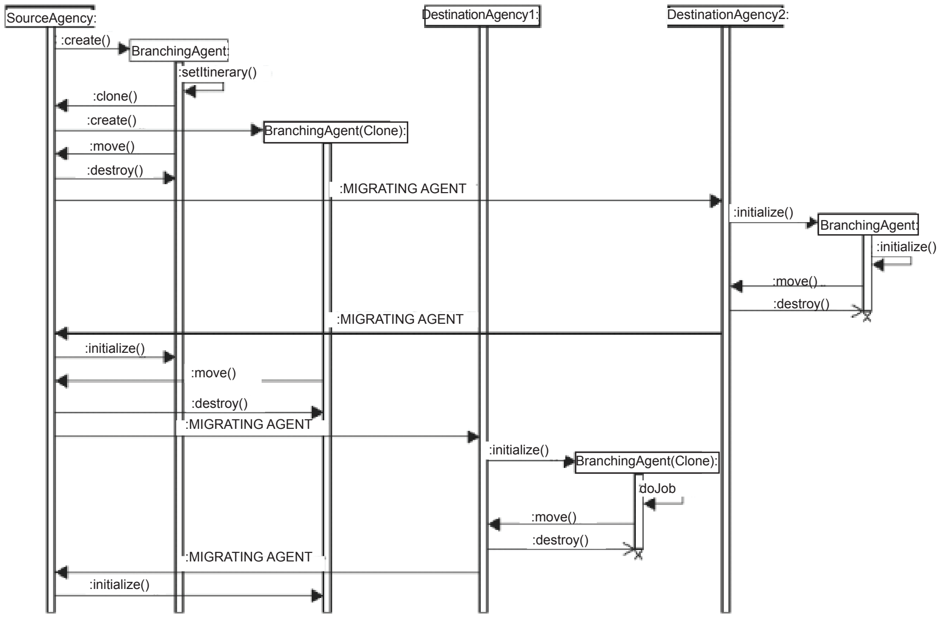
Figure 2. Message sequence chart for Star-Shaped


SourceAgency:
DestinationAgency1:
DestinationAgency2:
:create()
StarShapedAgent:
:setItinerary()
:nextDestination()
:move()
:destroy()
:MIGRATING AGENT
:initialize()
StarShapedAgent:
doJob
:move()
:destroy()
:MIGRATING AGENT
:initialize()
:showResults()
:nextDestination()
:move()
:destroy()
:MIGRATING AGENT
:initialize()
StarShapedAgent:
doJob
:move()
:destroy()
:MIGRATING AGENT
:initialize()
:showResults()

the first destination agency in the list. After migration is completed, it executes the relevant job and resumes migration going back to the source agency. The agent repeats this cycle until the last agency on its list is visited. The advantage of this pattern is that the agent stores the results of its job in the source agency and does not need to migrate to the others' agency with them. Depending on the application, the results can be shown to the user as soon as the agent stores them in the source agency. In this way, the user can already know the partial results before the agent finishes its migration through all search agencies.

In Figure 2, we can see an execution sequence for the Star-Shaped pattern. In this diagram, we have the same configuration of the sequence diagram shown for the itinerary pattern: three agencies and one agent. Following the diagram, we observe that the agent sets its itinerary and then travels to the first search agency. After executing its job, the agent returns to the source agency, where it stores the job's result.

Figure 3. Message sequence chart for branching



After that, the agent travels to the second search agency, executes its jobs, and returns to the source agency, storing the results obtained.

Branching

In the branching pattern, the agent receives a list of agencies to visit and clones itself according to the numbers of agencies in the defined itinerary. Each clone is assigned an agency from the received list. Each clone has to migrate to its corresponding agency, execute its job, and notify the source agency when the job is completed. The importance of this pattern is that it splits the tasks that can be executed in parallel. The treatment of the final results is an issue not covered by this pattern. For instance, the clones can put the result of the task in a user interface or send it to another agent.

Figure 3 shows an execution sequence for this pattern, in a scenario where there are three agencies and one agent. Following this figure, we can see that the agent sets its itinerary and then clones itself. After that, each agent (the original and the clone) migrates to a search agency, where they execute the job and then return to the source agency.

Petri Nets

A colored Petri net (Jensen, 1992, 1997), or simply CPN, is a formal method with a mathematical base and a graphical notation for the specification and analysis of systems with characteristics such as concurrent, parallel, distributed, asynchronous, timed, among others. For the mathematical definition, the reader can refer to Jensen (1992, 1997). The graphical notation is a bipartite graph with places, represented as ellipses, and transitions, represented as rectangles. Transitions represent actions, and the marking of the places represents the state of the model. A marking of a place at a given moment is the token present at that place. A token can be a complex data type in CPN/ML language (Christensen & Haagh, 1996). Each place has an associated color set that represents the kind of tokens the place can have. The transitions can have guards and code associated to it. Guard is a Boolean expression that must be true for the transition to fire. Code can be a function that is executed every time the transition fires. Arcs go from places to transitions and from transitions to places, and never from transition to transition or place to place; they can have complicated expressions and function calls associated to them.

For a transition to fire, it is necessary that all input places, that is, places that have arcs that go from the place to the transition, have the number of tokens greater than or equal to the weight of the arc, w(p,t), and the guard of the transition must be true. When these characteristics hold the transition is said to be enabled to fire. An enabled transition can fire at any time and not necessarily immediately. Once a transition fires it removes w(pi,t) from each input place pi, and the output places, that is, the places that have arcs from the transition to the place, receive tokens according to the arc expression from the transition to the place: w(t,p).

For a detailed explanation of colored Petri nets, consult Jensen (1992, 1997).

Object-oriented Petri net (RPOO) models consist of a set of classes and their corresponding Petri nets. Classes are described and can be related to each other like UML classes. The Petri nets describe the internal behavior of objects. For each object, there is one Petri net that models it. A variety of RPOO actions (instantiate objects, call methods, destroy objects, etc.) that can be performed by the objects and actions are depicted by transition inscriptions in the colored Petri nets. An inscription may describe several actions, and all the actions in an inscription are executed atomically. A set of atomically executed actions is called an event.

In RPOO, each object is a thread, and interaction between two objects may be asynchronous. This means that when an object *a* calls a method of (sends a message to) object *b* in asynchronous mode, the system moves to a state where the data passed as parameter will be pending and may be consumed in a further action by object *b*. RPOO actions also include synchronous calls, where messages are sent and consumed atomically.

A set of interconnected objects and its pending messages form a structure of an object system. Besides this structure, an object system also knows what is called imminent actions, that is, actions that may be executed in the current structure. Briefly, an action will be executed, in case of concurrency.

In RPOO, synchronously sent messages are presented by inscriptions like obj.set(data). In other words, the current object is sending a message to obj object. To send a synchronous message, the exclamation point obj!set(data) is used. To fire a transition when a message has been waiting, there must be an inscription like obj?set(data), denoting that a message is waiting.

For a more detailed explanation of RPOO, consult Guerrero (2002).

FORMAL MODELS AND ANALYSIS OF PATTERNS

Before modeling the patterns, it is necessary to identify all the entities that will be part of the whole model. And it is also necessary to build a diagram class containing these entities and their relationships to each other.

Figure 4 shows the class diagram for this system. There are three types of agents, one for each one of the migration patterns presented in this article: ItineraryAgent, StarShapedAgent, and BranchingAgent. An agent has a list of agencies that represents its destination. At the beginning of the model, the current agent belongs to an agency, the source one. All the relationships change during the execution of the model. The class diagram represents the initial state of the scenario.

RPOO Models

There is no tool for editing RPOO models, so the famous tool set Design/CPN (CPN Group, 1999) is used to edit and analyze such models. Each class is declared using the CPN/ML inscriptions (Christensen & Haagh, 1996).

Figure 4. Class diagram of design patterns

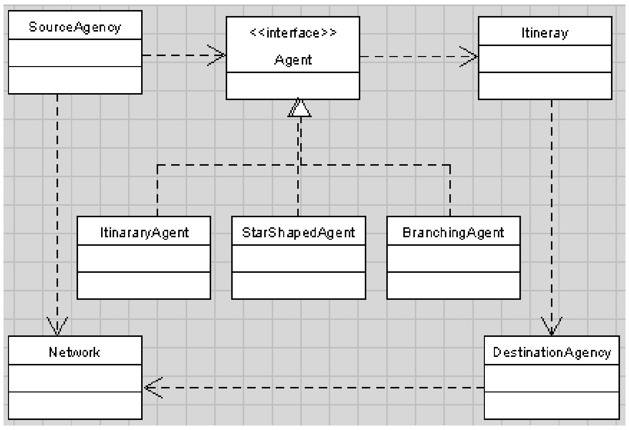


Figure 5 shows the model for the Itinerary class. In this model, we see the behavior of the class and its relationship with agent classes that implement agent interface and with SourceAgency class.

Figure 6 shows the modeling of the DestinationAgency class. There is no complex behavior here; just the initialization of the agent is modeled. This initialization also represents the jobs to be executed locally.

In Figure 7, the SourceAgency class is modeled. Its actions, based on its relationships with itinerary, network, and BranchingAgent, are presented.

Figure 8 shows an abstraction of the network (e.g., Internet). It models the network class. Agents and agencies are transmitted through the network. Each agency receives its corresponding agent.

Figure 9 shows the RPOO model for the ItineraryAgent class. It models the Itinerary migration pattern. The agent goes to every destination agency, it does its local job, and it returns to the source agency.

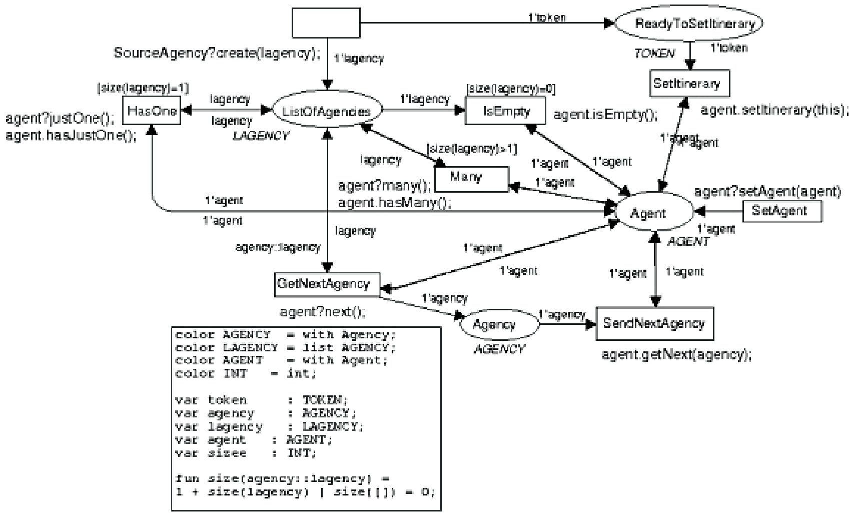
The StarShapedAgent class is modeled by the RPOO net presented in the Figure 10. This model interleaves each destination agency with the source agency. With this, the agent always visits its source agency after it makes a job remotely.

And Figure 11 presents the RPOO model for the BranchingAgent class. For each different destination agency, the model shown in Figure 7 makes a clone of this agent. Each one of these clones receives a list of destination agencies containing just one agency. Each one of the clone agents goes to the remote agency, does its job locally, and then returns with the results to the source agency.

Analysis

For each one of the mobile migration patterns presented here, simulations were done using the classical CPN simulator

Figure 5. RPOO model for Itinerary class

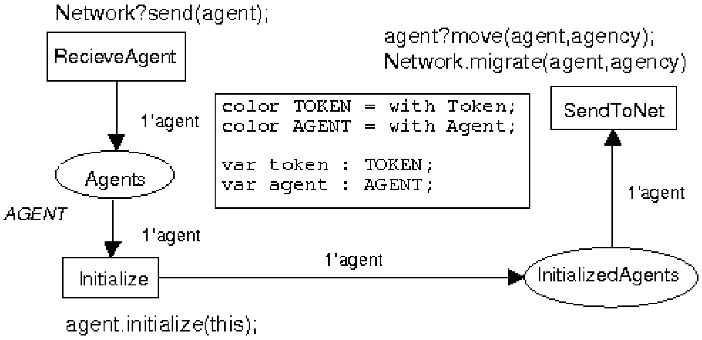

1'token
ReadyToSetItinerary
TOKEN
1'token
SetItinerary
agent.setItinerary(this);
SourceAgency?create(lagency);
1'lagency
[size(lagency)=1]
HasOne
agent?justOne();
agent.hasJustOne();
lagency
lagency
LAGENCY
ListOfAgencies
1'lagency
[size(lagency)=0]
IsEmpty
agent.isEmpty();
1'agent
1'agent
[size(lagency)>1]
lagency
Many
1'agent
1'agent
1'agent
agent?many();
agent.hasMany();
agent?setAgent(agent)
SetAgent
1'agent
1'agent
Agent
1'agent
AGENT
lagency
agency::lagency
1'agent
1'agent
1'agent
1'agent
GetNextAgency
agent?next();
1'agency
Agency
AGENCY
1'agency
SendNextAgency
agent.getNext(agency);

```
color AGENCY  = with Agency;
color LAGENCY = list AGENCY;
color AGENT   = with Agent;
color INT   = int;

var token     : TOKEN;
var agency    : AGENCY;
var lagency   : LAGENCY;
var agent   : AGENT;
var sizee   : INT;

fun size(agency::lagency) =
1 + size(lagency) | size([]) = 0;
```

Figure 6. RPOO model for DestinationAgency class


Network?send(agent);
RecieveAgent
1'agent
Agents
AGENT
1'agent
Initialize
agent.initialize(this);
1'agent
InitializedAgents
1'agent
SendToNet
agent?move(agent,agency);
Network.migrate(agent,agency)

```
color TOKEN = with Token;
color AGENT = with Agent;

var token : TOKEN;
var agent : AGENT;
```

Design/CPN. All the models behaved as expected. However, not all the possible behaviors were analyzed because of restrictions of existent tools. However, the RPOO models presented in this article showed how objected-oriented and mobile systems can be formally modeled and analyzed. To analyze all possible behavior, it is necessary to generate the state space of the formal model, also known as occurrence graph. With this graph and a model checker that supports RPOO model format, it will be possible to reason about some interesting properties described using some temporal logic.

BRIEF COMPARISON WITH CLASSICAL CPN MODELS

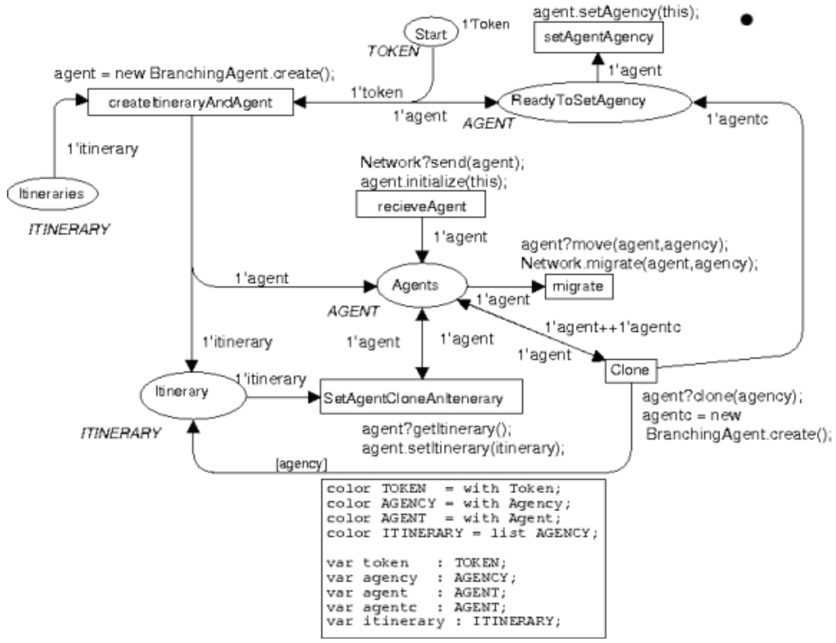
The models presented in this article have an object-oriented feature. This feature makes the modeling task easier because the formalism used is semantically closer to the modeled system. The classical colored Petri nets presented in Lima (2004) and in Lima, Figueiredo, and Guerrero (2004) do not have this feature, which makes the mapping between the Petri net world and the programming language work more difficult.

Due to this object-oriented feature, models can express the system behavior in a separate way. As it was said, each class in the class diagram is modeled in a unique RPOO model. This aspect makes all the modeled systems more organized and easier to be analyzed and studied.

FUTURE TRENDS

One future trend is to use RPOOt (Guerra, 2005; Guerra, Figueiredo, & Guerrero, 2005), or timed RPOO, to analyze the performance of each pattern presented here in some different contexts. With this, it will be possible to say more about them.

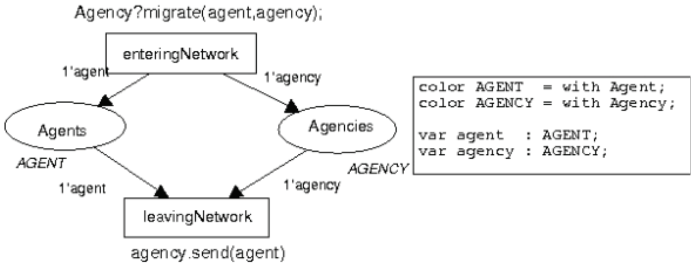
Figure 7. RPOO model for SourceAgency class


agent.setAgency(this);
setAgentAgency
Start
1'Token
TOKEN
1'agent
agent = new BranchingAgent.create();
createItineraryAndAgent
1'token
1'agent
ReadyToSetAgency
AGENT
1'agentc
1'itinerary
Itineraries
ITINERARY
Network?send(agent);
agent.initialize(this);
recieveAgent
1'agent
agent?move(agent,agency);
Network.migrate(agent,agency);
1'agent
Agents
migrate
AGENT
1'agent
1'agent++1'agentc
1'itinerary
1'agent
1'agent
1'agent
Clone
1'itinerary
Itinerary
SetAgentCloneAnItenerary
ITINERARY
agent?getItinerary();
agent.setItinerary(itinerary);
agent?clone(agency);
agentc = new
BranchingAgent.create();
[agency]

```
color TOKEN  = with Token;
color AGENCY = with Agency;
color AGENT  = with Agent;
color ITINERARY = list AGENCY;

var token   : TOKEN;
var agency  : AGENCY;
var agent   : AGENT;
var agentc  : AGENT;
var itinerary : ITINERARY;
```

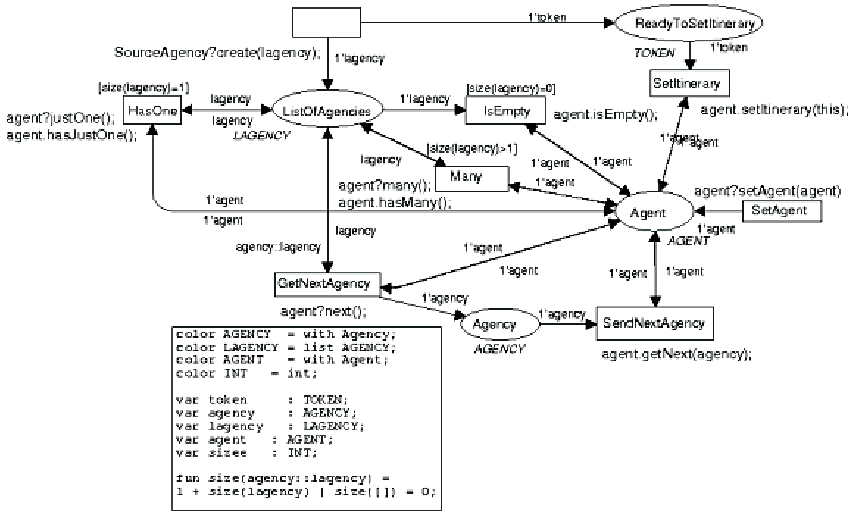
Figure 8. RPOO model for Network class


Agency?migrate(agent,agency);
enteringNetwork
1'agent
1'agency
Agents
AGENT
Agencies
AGENCY
1'agent
1'agency
leavingNetwork
agency.send(agent)

```
color AGENT  = with Agent;
color AGENCY = with Agency;

var agent  : AGENT;
var agency : AGENCY;
```

Figure 9. RPOO model for ItineraryAgent class


1'token
ReadyToSetItinerary
SourceAgency?create(lagency);
1'lagency
TOKEN
1'token
SetItinerary
[size(lagency)=1]
HasOne
lagency
lagency
ListOfAgencies
LAGENCY
1'lagency
[size(lagency)=0]
IsEmpty
agent.isEmpty();
agent.setItinerary(this);
agent?justOne();
agent.hasJustOne();
lagency
[size(lagency)>1]
Many
agent?many();
agent.hasMany();
1'agent
1'agent
1'agent
1'agent
1'agent
1'agent
1'agent
Agent
AGENT
agent?setAgent(agent)
SetAgent
1'agent
lagency
agency::lagency
1'agent
1'agent
1'agent
1'agent
GetNextAgency
1'agency
agent?next();
Agency
AGENCY
1'agency
SendNextAgency
agent.getNext(agency);

```
color AGENCY   = with Agency;
color LAGENCY  = list AGENCY;
color AGENT    = with Agent;
color INT   = int;

var token      : TOKEN;
var agency     : AGENCY;
var lagency    : LAGENCY;
var agent    : AGENT;
var sizee    : INT;

fun size(agency::lagency) =
1 + size(lagency) | size([]) = 0;
```

Figure 10. RPOO model for StarShapedAgent class

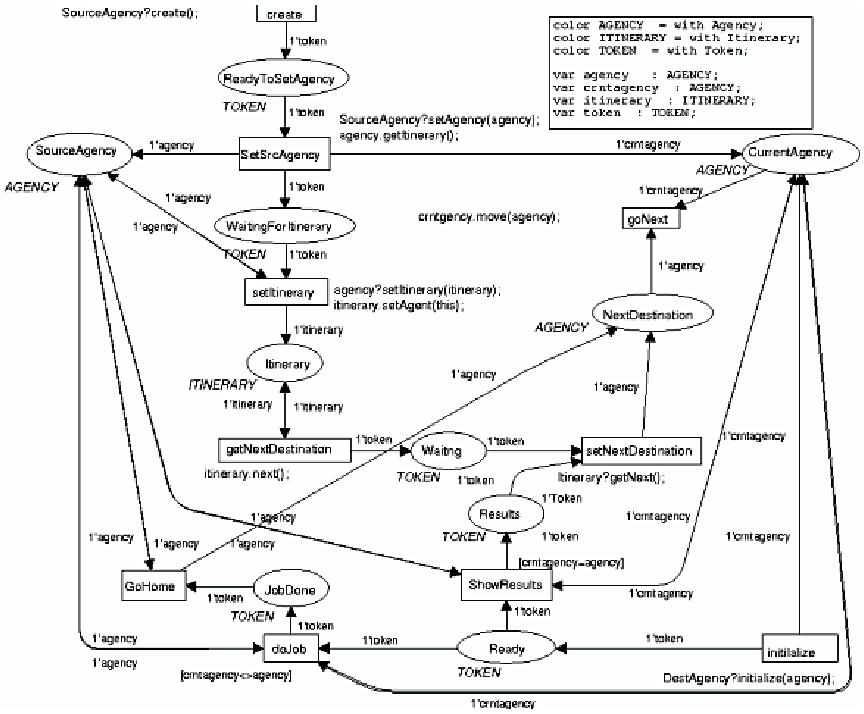
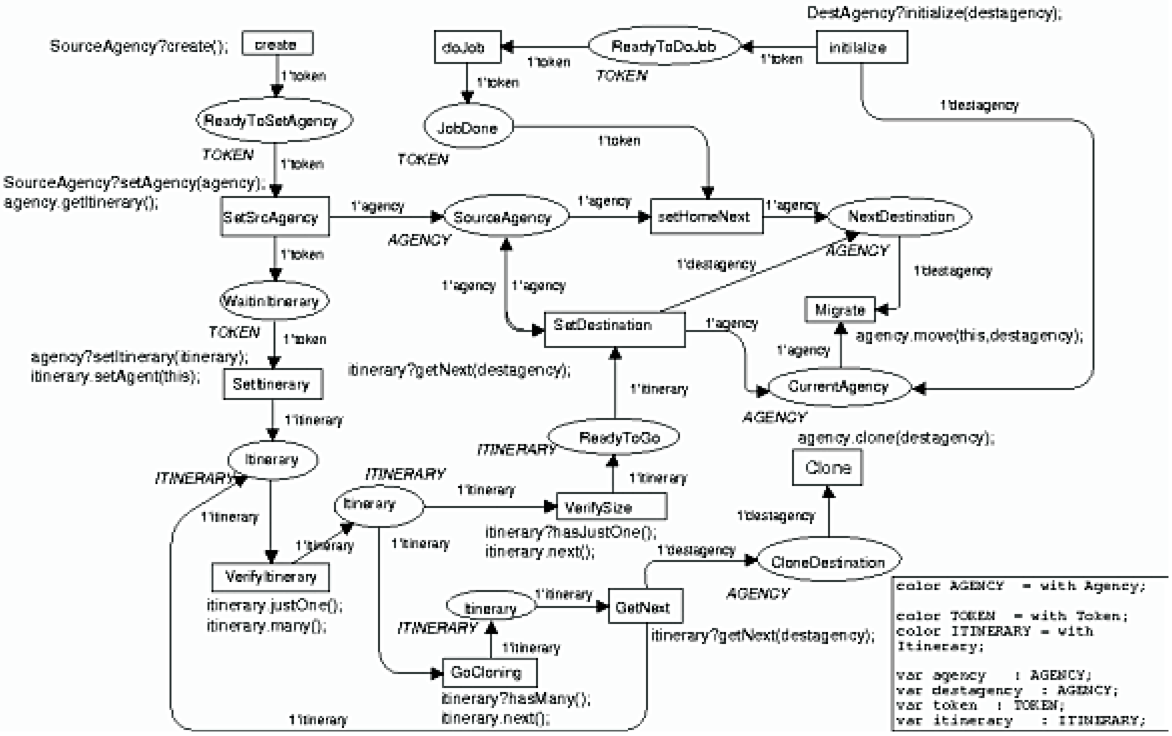

SourceAgency?create();
create
1'token
ReadyToSetAgency
TOKEN
SourceAgency?setAgency(agency);
agency.getItinerary();
SourceAgency
AGENCY
SetSrcAgency
1'agency
1'crntagency
CurrentAgency
AGENCY
WaitingForItinerary
TOKEN
crntgency.move(agency);
goNext
setItinerary
agency?setItinerary(itinerary);
itinerary.setAgent(this);
NextDestination
AGENCY
1'itinerary
Itinerary
ITINERARY
getNextDestination
itinerary.next();
Waitng
TOKEN
setNextDestination
Itinerary?getNext();
Results
TOKEN
1'Token
[crntagency=agency]
GoHome
JobDone
TOKEN
ShowResults
doJob
Ready
TOKEN
initialize
[crntagency<>agency]
DestAgency?initialize(agency);
color AGENCY = with Agency;
color ITINERARY = with Itinerary;
color TOKEN = with Token;
var agency : AGENCY;
var crntagency : AGENCY;
var itinerary : ITINERARY;
var token : TOKEN;

Figure 11. RPOO model for BranchingAgent class


DestAgency?initialize(destagency);
SourceAgency?create();
create
doJob
ReadyToDoJob
TOKEN
initialize
1'token
1'destagency
ReadyToSetAgency
TOKEN
JobDone
TOKEN
SourceAgency?setAgency(agency);
agency.getItinerary();
SetSrcAgency
1'agency
SourceAgency
AGENCY
setHomeNext
NextDestination
AGENCY
WaitinItinerary
TOKEN
SetDestination
Migrate
agency.move(this,destagency);
agency?setItinerary(itinerary);
itinerary.setAgent(this);
SetItinerary
itinerary?getNext(destagency);
CurrentAgency
AGENCY
1'itinerary
ReadyToGo
ITINERARY
agency.clone(destagency);
Clone
Itinerary
ITINERARY
VerifySize
itinerary?hasJustOne();
itinerary.next();
CloneDestination
AGENCY
VerifyItinerary
itinerary.justOne();
itinerary.many();
GetNext
itinerary?getNext(destagency);
GoCloning
itinerary?hasMany();
itinerary.next();
color AGENCY = with Agency;
color TOKEN = with Token;
color ITINERARY = with Itinerary;
var agency : AGENCY;
var destagency : AGENCY;
var token : TOKEN;
var itinerary : ITINERARY;

Besides, only a few simulations were done with the models presented above. As is known, simulations do not cover all possible behaviors of the model. As mentioned before, to do this it is necessary to generate the occurrence graph. The state space generator for RPOO presented in Silva (2005) will be used to obtain the occurrence graph. It will check for some of the properties described in CTL (Sifakis, 1990). To reason about these temporal properties, the model checker presented in Rodrigues (2004) will also be used.

CONCLUSION

As it can be seen, modeling a system with an object-oriented Petri net is not a trivial task. However, the benefits it brings to the development process is worth the effort. RPOO has an advantage over classical colored Petri nets: their semantics are closer to the programming language paradigm. But it is clear that the absence of a specific tool for editing RPOO models represents a large disadvantage for its usage for verifying and analyzing tasks.

It is clear that RPOO is a good formalism to model mobile systems. Due to its object-oriented feature, the whole model is kept organized. Based on this experiment, RPOO showed to be more interesting in modeling such a mobile system than the classical colored Petri nets. But, a larger experiment is needed to show if this feature is general or if it was just in this specific context.

REFERENCES

Christensen, S., & Haagh, T. B. (1996). *Design/CPN overview of CPN ml syntax.*

CPN Group (1999). *Design/CPN 4.0.* University of Aarhus, Denmark. Retrieved from http://www.daimi.au.dk/designCPN/

Guedes, F. P., Machado, P. D. L., & Medeiros, V. N. (2003). *Developing mobile agent based applications.* La Paz: Latin American Center of Studies in Computer Science.

Guerra, F. V. de A. (2005). *Modelagem de sistemas com restrições temporais em Redes de Petri orientadas a objetos.* Master thesis, Curso de Pós-Graduação em Informática, Universidade Federal de Campina Grande, Brasil.

Guerra, F. V. de A., Figueiredo, J. C. A. de, & Guerrero, D.S. (2005). Protocol performance analysis using a timed extension for an object oriented Petri net language. *Electronic Notes on Theorical Computer Science,* (130), 187-209.

Guerrero, D. D. S. (2002). *Redes de Petri orientadas a objetos.* PhD Thesis, Curso de Pós-Graduação em Engenharia Elétrica, Universidade Federal da Paraíba–Campus II, Brasil.

Jensen, K. (1992). *Coloured Petri nets: Basic concepts, analysis, methods and practical use.* Berlin: Springer-Verlag.

Jensen, K. (1997). *Coloured Petri nets: Basic concepts, analysis methods and practical use* (vol. 2). Berlin: Springer-Verlag.

Klein, C., Rausch, A., Sihling, M., & Wen, Z. (2001). Extension of the Unified Modeling Language for mobile agents. In K. Siau & T. Halpin (Eds.), *Unified Modeling Language: Systems analysis, design and development issues* (pp. 116-128). Hershey, PA: Idea Group Publishing.

Lima, E. F. de A. (2004). *Formalização e análise de padrões de projeto para agentes móveis.* Master's Thesis, Curso de Pós-Graduação em Informática, Universidade Federal de Campina Grande, Brasil.

Lima, E. F. de A., Figueiredo, J. C. A. de, & Guerrero, D. S. (2004). Using coloured Petri nets to compare mobile agent design patterns. *Electronic Notes on Theorical Computer Science,* (95), 287-305.

Medcraft, P. S. (2003). *Integração de bancos de dados federados na Web usando agentes movies.* Master's Thesis, Universidade Federal de Campina Grande, Brasil.

Rodrigues, C. L. (2004). *Verificação de modelos RPOO.* Master thesis, Curso de Pós-Graduação em Informática, Universidade Federal de Campina Grande, Brasil.

Sifakis, J. (Ed.). (1990). *Proceedings of the International Workshop on Automatic Verification Methods for Finite State Systems.* Berlin: Springer-Verlag (LNCS 407).

Silva, T. M. de (2005). *Simulação automática e geração de espaço de estados de modelos em Redes de Petri orientadas a objetos.* Master thesis, Curso de Pós-Graduação em Informática, Universidade Federal de Campina Grande, Brasil.

Tahara, Y., Ohsuga, A., & Honiden, S. (1999). Agent system development method based on agent patterns. *Proceedings of the 21st International Conference on Software Engineering* (pp. 356-367). Los Angeles, CA.

KEY TERMS

Agent Migration Pattern: A solution for a given problem inside the context of mobile agents.

Computation Tree Logic (CTL): A type of temporal logic in which different futures are considered.

Mobile Agent: A piece of computer software that is able to migrate (move) from one computer to another autonomously and continue its execution on the destination computer.

Object-Oriented Petri Net: A formalism that puts together the Petri net formalism and the object-oriented paradigm. In Portuguese: Redes de Petri Orientadas a Objetos (RPOO).

Occurrence Graph: A graph in which each node represents a single possible state of the model.

RPOO (Redes de Petri Orientadas a Objetos): See *Object-Oriented Petri Net.*

RPOOt (Redes de Petri Orientadas a Objetos temporizadas): A timed *Object-Oriented Petri Net.*

Temporal Logic: Term used to describe any system of rules and symbolism for representing, and reasoning about, propositions qualified in terms of time.

Virtualization and Mobility in Client and Server Environments

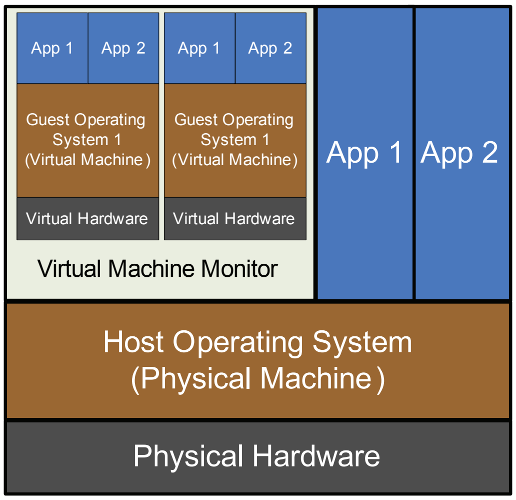
Eduardo Correia
Christchurch Polytechnic Institute of Technology, New Zealand

INTRODUCTION

A great deal of popular software is not designed for mobility (Griffiths, 2004). This is peculiar because many mobile users expect to have easy access to an information infrastructure that links up their mobile phone, laptop, personal digital assistant (PDA), and other devices, while the backend systems of organizations need to be agile, especially as the number, range, and diversity of services and associated technologies grow. Enter virtualization, a technology that has been part of computing for many years, but only fairly recently become mainstream (Intel, 2006; Singh, 2004). It makes use of a virtual machine monitor (VMM), a mechanism that frees up systems from many of the physical constraints of hardware, by adding a software layer that abstracts hardware, so that an entire machine, operating system, applications, and even data can be stored as a set of standard folders and files. While it is well established that this architecture enhances security and reliability (Rosenblum & Garfinkal, 2004), it also enables both users and systems, as this article shows, to be mobile and responsive to change, both in client and server environments.

VMware Workstation, Microsoft Virtual PC, and other virtualization software takes the form of a standard application that can be installed on physical computers. As Figure 1 shows, these applications provide a VMM, which enables one system (the guest) to run within the context of another system (the host). The VMM presents a complete set of virtual hardware to each guest virtual machine (VM) running within this environment. Just as ordinary computers access physical resources, such as memory, processors, hard disks, and network adapters, so too do each of the virtual or guest systems, only their hardware is an instance of a generic abstraction that the VMM generates for each of them. The VMM then mediates various calls made by VMs to access the physical hardware of the host machine. Whereas the standard computer has a single operating system with applications installed to it, a computer with virtualization software runs, in addition, within the VMM one or more operating systems, each with their own applications installed.

Figure 1. Virtualization architecture


App 1
App 2
App 1
App 2
Guest Operating System 1 (Virtual Machine)
Guest Operating System 1 (Virtual Machine)
Virtual Hardware
Virtual Hardware
Virtual Machine Monitor
App 1
App 2
Host Operating System (Physical Machine)
Physical Hardware

MOBILE VIRTUAL MACHINES

While virtual private networks enable users to work from remote locations, as if they are sitting at a machine on the local network to some extent, this approach has certain drawbacks. Users may wish to connect with machines that do not belong to the organization and therefore do not adhere to its policies and standards. Antivirus software may be out of date or updates not installed, for example. One solution is to provide a quarantine area that will allow a machine into the network, but restrict its access to resources until certain criteria have been met. Cisco Systems' Network Admission Control (NAC) and Microsoft's Network Access Protection (NAP) are examples of this kind of solution (Conry-Murray, 2005). Alternatively, network administrators can make use of the VMware Assured Computing Environment (ACE) to produce and deploy secure, fully built virtual machines that apply custom policies and adhere to certain specified standards (Burt, 2004).

The fact that it is in effect the VM that forms part of the network and not the physical machine means that it does not matter to the network administrators that these particular hosts may not have the latest antivirus signature files or applied recent updates, as this underlying (physical) system does not interact with the network and cannot influence it in any way. Naturally, the user's physical machine could fail causing the VM itself to fail, but this will still not affect the network, and restoring the client machine is simply a matter

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

of copying the ACE VM from removable media, such as DVD or a pen drive (VMware, 2006a). In this way anyone needing temporary access to the network can be given it because perhaps the only way of accessing it using their own machine is through a VM compliant with the standards set by the organization, including exactly which resources can be accessed and the length of time the VM can be used.

SYSTEM AND SESSION MIGRATIONS

When conventional systems need to be moved from one set of hardware to another, the operating system and server applications first need to be installed and then the data, assuming it is on the same system, moved, either by simply copying it or restoring it from a backup. It is not just that this is a time-consuming exercise, but also it often entails having to navigate the complexities of making a system work with significantly different hardware, and its associated drivers and other software. This can cause lengthy outages and make system migrations complex and arduous. Virtual systems, by contrast, always access the same set of physical resources the VMM presents to it, whatever the differences in the actual underlying physical hardware, making them much more mobile than conventional systems. In fact, moving the entire system is simply a matter of copying data from one (physical) machine to another, making VMs highly portable (Wolf & Halter, 2005, p. 481). With Vmotion, systems administrators can even move virtual machines without interrupting service availability (VMware, 2006b), something that has been used with success in live production environments (Rosenkoetter, 2006). This makes it feasible to apply many changes during business hours that would otherwise have been scheduled for weekends or at least when the network is quiet (Cline, 2006).

In such cases virtualization enables easier migration of systems from one set of hardware to another, and significantly reduces the risk associated with such change. This risk is reduced further by the ease with which it is possible to store multiple older versions of virtual systems. Where for instance poor software or problematic devices are installed, virtualization enables administrators to return the machine easily to a selected previous state. This concept can be taken a step further by capturing and virtualizing a client's entire session with an existing server, then migrating to another system that is the same or different, such as from a desktop machine to a PDA (Baratto, Potter, Su, & Nieh, 2004). In fact virtualization can even be used by mobile users to "decouple" a computer into a "body (display, CPU, RAM, I/0) and a soul (session state, software, data, preferences)" so that a user with a portable device can walk up to any computer and resume a session started on another machine (Cáceres, Carter, Narayanaswami, & Raghunath, 2005, p. 65).

CONCLUSION

Virtualization has expanded dramatically in recent years because it is a flexible, scalable technology. It allows powerful hardware to be used more efficiently by distributing the processing, storage, and movement of data among several virtual systems that can still make use of clustering and other conventional forms of load balancing and redundancy. VMware ESX Server and Microsoft Virtual Server for instance make it cost effective to host a single major application per server, so reducing software conflicts and increasing reliability by isolating each of the guest systems, as well as the host from one another while enabling systems administrators to fine tune systems to resident applications. It also makes it easy to retain (copies of) entire systems either for the purposes of disaster recovery or to test future development, perhaps with a view to implementing such changes on production systems.

Virtualization can be utilized in a range of situations in both client and server environments, be it as part of a mission-critical system users connect to; a disaster-recovery infrastructure; a gateway for connecting securely to the network; a deployment of secure, fully built clients that comply with specified standards; or a portable learning environment comprising an entire network of virtual machines that may or may not be connected to virtual and even physical switches, but which can be easily moved from one physical machine to another. It is true that virtualization often demands powerful hardware in order for it to cope with the demands of hosting numerous systems; but with the reduction in the cost of hardware, the use of virtual systems becomes a viable proposition for organizations, especially as they are easier to manage and more agile than conventional systems. It is no wonder then that manufacturers are beginning to produce hardware that is designed with virtualization in mind (Intel, 2006), and that VMware and Virtual PC have become such well-known brands in recent times. According to one survey, respondents expected 45% of new servers deployed this year to make use of virtual machines (IDC, 2005), a trend, it appears, that is set to continue.

REFERENCES

Baratto, R. A., Potter, S., Su, G., & Nieh, J. (2004, September 26-October 1). MobiDesk: Mobile virtual desktop computing. *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking,* Philadelphia.

Burt, J. (2004, September 20). *VMware takes virtual machines mobile.* Retrieved April 5, 2006, from http://www.eweek.com/article2/0,1895,1647632,00.asp

Cáceres, R., Carter, C., Narayanaswami, C., & Raghunath, M. (2005, June 6-8). Reincarnating PCs with portable SoulPads. *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, Applications, and Services* 9pp. 65-78), Seattle, WA.

Conry-Murray, A. (2005). *Caymas Systems' access gateways.* Retrieved May 10, 2006, from http://www.itarchitect.com/shared/printableArticle.jhtml?articleID=171000823

Cline, K. (2006, March 1). *Re: Interesting comment from MS tech specialist about VS & ESX.* Posted on VMTN Discussion Forums; retrieved from http://www.vmware.com/community/thread.jspa?threadID=33012&start=30&tstart=0

Griffiths, G. (2004). *Help for the mobile worker: Keeping your remote workforce productive and secure.* Retrieved May 7, 2006, from http://www.everdream.com/wp/mobile_workforce_whitepaper.asp

IDC. (2005). *Increasing the load: Virtualization moves beyond proof of concept in the volume server market, according to IDC.* Retrieved May 7, 2006, from http://www.idc.com/getdoc.jsp?containerId=prUS00259905

Intel. (2006). *Intel virtualization technology: Hardware-assisted virtualization for today's businesses.* Retrieved May 7, 2006, from http://www.intel.com/products/processor/xeon/vt_prodbrief.pdf

Rosenblum, R., & Garfinkal, T. (2005). Virtual machine monitors: Current technology and future trends. *Computer, 38*(5), 39-47.

Rosenkoetter, R. (2006, March 1). *Re: Interesting comment from MS tech specialist about VS & ESX.* Posted on VMTN Discussion Forums; retrieved from http://www.vmware.com/community/thread.jspa?threadID=33012&start=30&tstart=0

Singh, A. (2004). *An introduction to virtualization.* Retrieved May 9, 2006, from http://www.kernelthread.com/publications/virtualization/

VMware. (2006a). *VMware ACE facilitates and streamlines remote working at Siemens Industrial Turbomachinery.* Retrieved April 5, 2006, from http://www.vmware.com/customers/stories/siemens_ace.html

VMware. (2006b). *VMware VMotion: Delivering game changing virtual machine mobility.* Retrieved May 9, 2006, from http://www.vmware.com/products/vc/vmotion.html

Wolf, C., & Halter, E.M. (2005). *Virtualization: From the desktop to the enterprise.* Berkeley, CA: Apress.

KEY TERMS

Application: A program that performs a specific task or related group of tasks and which requires an operating system to run successfully.

Guest: A virtual machine that runs in the context of a virtual machine monitor, which provides for it an abstracted hardware environment. See *Virtual Machine.*

Host: A node on a network that requires an IP address to communicate with other hosts, but in this article refers specifically to a physical computer that runs a virtual machine monitor capable of running one or more virtual machines.

Migration: The movement of operating systems, applications, and data from one computer to another; it can now also include the transfer of a session from one device to another.

Operating System: A general-purpose program that performs many basic tasks such as accepting input from the keyboard or printing output to a screen, as well as mediating access to software and hardware resources.

Virtualization: Partitioning a physical machine into a number of virtual machines, each of which effectively functions the same as a physical machine and can replace a physical machine on a network.

Virtual Machine: A machine much like a physical machine, in that it requires an operating system, can have applications installed, can present itself in exactly the same way as a physical machine on the network, but which functions within the specific abstracted hardware environment of a virtual machine monitor.

Virtual Machine Monitor: An environment created by an application, such as VMware or Virtual PC, with the purpose of enabling one or more virtual machines to run.

Voice Recognition Intelligent Agents Technology

Călin Gurău
Montpellier Business School, France

INTRODUCTION

Mobile computing technology is evolving at a rapid pace. Under the pressure of market demands, the format of mobile devices evolves towards a contradictory situation: on one hand, the handset tends to become smaller, but on the other hand, the users demand increased data search, transmission, and saving capabilities. In order to achieve this, the model of interaction between humans and mobile devices has to evolve from the presently prevalent keyboard-screen system for data input and output towards *voice-recognition intelligent agents* technology.

This article attempts to present the rationale and the advantages of this development, and to analyze the possible problems raised by the introduction of this technology.

The article starts with a presentation of the existing mobile phone technology, outlining its main limitations in terms of functionality, which are logically determined by the way and the context in which mobile phones are normally used. Based on the analysis of the contradictions between the present model of interaction with mobile phones and the requirements of users, the article presents possible solutions to this problem. The study argues that the introduction of voice recognition intelligent agents can enhance significantly the functionality of mobile phones, representing a true revolution in mobile computing. Various practical applications of this technology are briefly presented, as well as the main problems related with its development and implementation.

The article ends with a summary of the arguments discussed and with definitions of the main terms and concepts presented.

BACKGROUND

Mobile phone technology was developed with the aim to provide users with a telephone connection anyplace, anytime. The main innovation that allowed the mass adoption and use of mobile phones was the cellular approach in transmitting a radio signal. Traditionally, people that required frequent communications could install in their car a radio telephone, but the small number of radio channels available in one area limited drastically the number of possible users of this technology. By dividing a large area into small cells, and each of these cells having a low-power transmitter, the number of communication channels increases significantly, since people that are not located in neighboring cells can use the same frequency to communicate (Layton, Brain, & Tyson, 2005).

The introduction of digital technology (2G) has increased even further the number of communication channels. Finally, 3G technology represents the latest trend in mobile phones standards, offering increased bandwidth and information transfer rates to accommodate Web-based applications and phone-based audio and video files.

However, the use of mobile phones to access Internet applications presents a number of limitations, some of which are related with the specific interface of mobile phones, and others with the existing Web protocols adapted for mobile networks. The screens of mobile phones are small and have a lower resolution in comparison with PC or laptop screens/monitors. On the other hand, the wireless application protocol (WAP) works badly on wireless devices with small screens, and it is dependent on mobile technology's bandwidth (such as GSM or CDMA) for access to information and services (Yeo, & Huang, 2003). Other problems are connected with the Web navigation and site structure, or with the input methods available for mobile phone users (Buchanan et al., 2001).

The future development of mobile services requires a revolution in the technological model applied, which is based on transforming the Web architecture, as well as the usage of mobile devices.

WIRELESS WEB APPLICATIONS FOR MOBILE PHONES

At the moment, the Internet is a network of databases supported by applications that allow users to search, retrieve, and use information contained in computers' memory. However, the rapid increase of online available data makes it more and more difficult for users to find the specific information they need. A trivial search on the Google search engine usually displays a list of a few million more or less relevant Web documents. In this case, the use of a search engine is only the beginning and not the end of searching for particular data,

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

and this process can indeed be very time consuming, without providing any guarantees for a successful result.

An alternative is to use customized intelligent agents, which can search the Web for clearly defined data. These intelligent agents are usually adapted for a specific type of Web search—for example, they can provide a list of companies that offer online a specific type of product. Strictly speaking, they are not very intelligent, because traditionally, these applications were not able to learn and improve in time their searching capabilities. However, this can be changed. Using neural network technology, and registering a history of operations realized or a particular user, the advanced intelligent agents can progressively learn the preferences of their customers and provide improved results.

Wireless devices such as mobile phones have a number of limitations determined by their specific circumstances of usage. The main advantage of a mobile phone is obviously its mobility, which implies a small size and weight, combined with good usability, in various environment and circumstances. These characteristics limit the size and the resolution of the screen, the size, and the functionality of keypads; the power and memory capacity; as well as the bandwidth. Because of these problems, the mobile phones cannot be used like a PC device, which incorporates autonomous computing capabilities. However, the model of distributed networks and resources can be effectively applied to a mobile phone (Mattern, 2000). In a traditional sense, mobile phones are simple communicating devices, more like interface terminals than personal computers. The solution for their effective use is to create easily accessible networks with distributed resources, and to develop advanced software applications in order to improve the mobile communication capability (Alesso, & Smith, 2001).

Some of these applications that can significantly improve the interface between the user and the distributed network are based on voice-recognition technology. On the other hand, the nature of mobile devices is adapted to a model of discontinuous, time-limited use; therefore, the user does not have the time himself/herself to browse the Internet in search of relevant information. Even given the required time and stability, the user might prefer the use of an Internet-connected PC or terminal, which offers better interaction and visualization capabilities. This problem can be solved by developing intelligent applications that can work automatically and independently of any human supervision, using the instructions given by the user.

The solution of improved Web services using mobile devices is the combination of voice-recognition technology with the use of intelligent agents (Lai, Mitchell, Viveros, Wood, & Lee, 2002). The mobile phone user will initiate a command by directly speaking with one or more intelligent agents, which then can search for specific information on the Web and announce the results through an SMS message. This type of interaction is presented in Figure 1.

The intelligent agents pass through a succession of phases in order to fulfill their tasks (Rodríguez, Favela, & Muñoz, 2003):

- **Activated:** This represents the main state of the agent, which includes a number of specific sub-states:
  - **Learning:** Represents the initial sub-state of an agent, in which the agent acquires knowledge

Figure 1. The use of voice recognition intelligent agents in mobile phones services

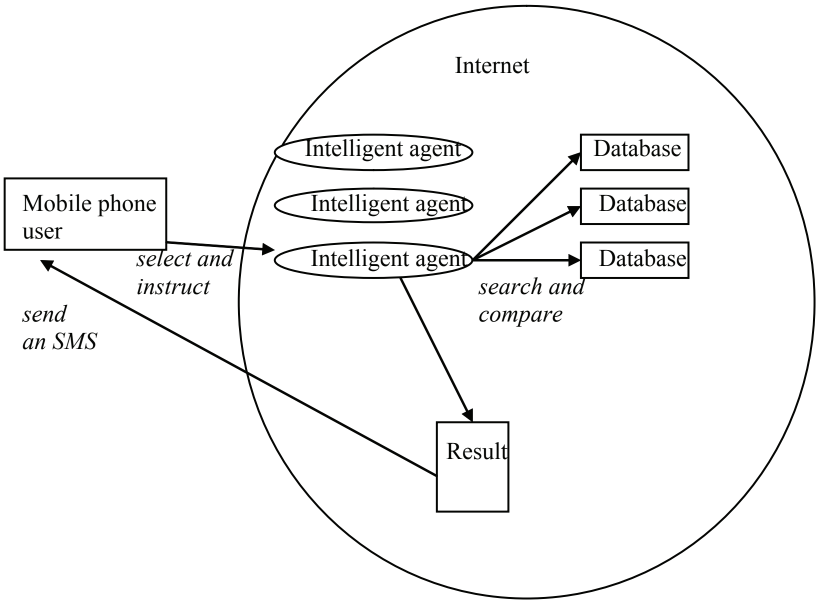
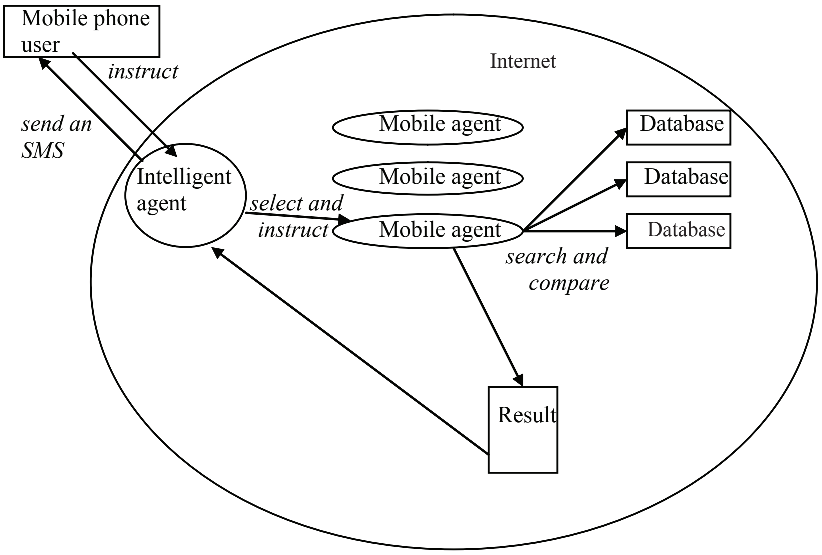


Figure 2. The use of voice recognition intelligent agents and mobile agents in mobile phones services



about its environment in various ways (through direct instructions from the user, a service, or other agents; from a sensor connected with the working environment; or through autonomous learning).

- **Analyzing:** On the basis of the data obtained in the previous sub-state, the agent establishes its goal and the method of fulfilling it.
- **Executing:** The agent performs the action plan established in the previous sub-state.
- **Communicating:** The agent can communicate with the user or with other agents, in order to collect additional information or to presents the results of its task.
- **Suspended:** While it waits for information, the agent can suspend its activity.

However, this model has two major problems:

- For the moment, the intelligent agents active on the Web are strictly specialized for specific tasks. The search for a specific agent can be cumbersome for the mobile phone user in terms of time and interface. In fact, in this case, the search for online information is replaced by the search for the best intelligent agent that can perform the automatic search.
- The size of software applications that correspond to Web intelligent agents is quite significant. If for the classical online navigation this is not a problem, the use of these applications in a mobile phone network—where bandwidth is still limited—can reduce the speed of the entire operation.

A model developed by Kowalczyk et al. (2003) provides a possible solution to these problems. The model includes the use of two different types of agents (see Figure 2):

- **Intelligent Agents:** These are personalized software applications that manage the information search needs of a particular mobile phone user; they are stationary agents that are not able to migrate to other platforms.
- **Mobile Agents:** These are of a smaller size, specialized in specific tasks, active on the Web, and have low demands on network connection, quality, and online time, because after their migration to a new platform or device, the online connection must not be maintained while they are working locally.

This solution is particularly adapted to a distributed network architecture, which represents the most suitable infrastructure for mobile devices; for example, the personalized intelligent agent can be located on the home or office-based PC of the user, the mobile phone representing just a communication device between various elements of this process (Lino, Tate, Siebra, & Chen-Burger, 2003).

Despite its advantages, the large-scale use of voice-recognition agents for mobile phones is likely to develop a number of problems (Wong & Starner, 2001):

1. **Unreliable Voice Recognition:** The voice recognition techniques that are adequate for an office environment might not be feasible in mobile environments, where the voice power is likely to vary and the ambient noises

might deteriorate the quality of sounds (D'Agostino, 2005).

2. **Privacy:** The use of voice-recognition agents on mobile devices raises the problem of privacy of the user, since s/he is transmitting the instructions to the intelligent agent with a loud, clear voice.
3. **Cognitive Load and Attention:** The capacity of the human brain to coordinate multiple tasks is limited, which has to be taken into account considering that often, the users of mobile devices are also engaged in other activities.
4. **Security of Mobile Intelligent Agents:** This is noted by van Eijk, Hamers, Klos, and Bargh (2002) and can be found in various environments.

The possible applications of intelligent agents in mobile and pervasive computing have determined the development of various platforms for developing agent and multi-agent systems. On the other hand, the problems raised by platform compatibility have encouraged the standardization of different systems. The result of these efforts was concretized in two internationally recognized standards (van Eijk et al., 2002): the Foundation for Intelligent Physical Agents (FIPA) and the Object Management Group (OMG). One of the best agent platforms for mobile devices is the JADE/LEAP platform, which was the result of the integration between a project funded by an EU grant for the development of an agent platform implemented in the JAVA software language (JADE), and a project called the Lightweight and Extensible Agent Platform, initiated by a consortium consisting of Motorola, ADAC, BT, Broadcom Eireann Research, Telecom Italia Lab, University of Parma, and Siemens.

FUTURE TRENDS

These models of mobile phone Web services are still in the development stage. However, the missing elements of the process are gradually developed and integrated in systems with multiple applications. At the basis of this model is the concept of the Semantic Web. The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the idea of having data on the Web defined and linked such that it can be used for more effective discovery, automation, integration, and reuse across various applications (Hendler, Berners-Lee, & Miller, 2002).

Semantic Web services can be used, among others, for intelligent mobile construction collaborations (Aziz, Anumba, Ruikar, Carrillo, & Bouchlaghem, 2004), participation in Web site auctions (Kowalczyk et al., 2003), and banking services or real estate transactions (Clareity, 2004).

CONCLUSION

This article has attempted to provide a general overview of the main problems in mobile communication and computing, and has presented a possible solution through the application of voice-recognition intelligent agents to small mobile devices.

The future of mobile intelligent agents based on voice-recognition technology is full of opportunities. Various public and private organizations are already developing the various elements of this system. However, the successful implementation of these applications has to take into account the following issues:

1. The use of mobile intelligent agents requires a completely restructured Internet system, which defines and links various information and databases using the semantic model.
2. The applicability of intelligent agents to small mobile devices depends on the implementation of pervasive computing environments, with networks of distributed resources in terms of memory, databases, services, and applications.
3. The capacity of mobile intelligent agents to migrate among various platform raises issues of compatibility and security.
4. The use of voice-recognition technology also creates challenges at social, ethical, and moral levels (e.g., the use of voice-recognition systems in social environments or personal privacy issues).

These problems require an answer that can integrate the informational, technological, social, and personal levels before the mobile intelligent agents and voice-recognition technology will become widespread commercial applications.

REFERENCES

Alesso, H. P., & Smith, C. F. (2001). *The intelligent wireless Web.* Boston: Addison-Wesley Professional.

Aziz, Z., Anumba, C., Ruikar, D., Carrillo, P., & Bouchlaghem, D. (2004) Semantic Web based services for intelligent mobile construction collaboration. *ITCon, 9,* 367-379.

Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G., & Pazzani, M. (2001). *Improving mobile Internet usability.* Retrieved December 2005 from http://www10.org/cdrom/papers/230/

Clareity. (2004). *The perfect computer interface for the real estate industry.* Retrieved December 2005 from http://www.callclareity.com/2004-VR.cfm

D'Agostino, D. (2005). Weak speech recognition leaves customers cold. *CIO Insight,* (December 29). Retrieved January 2006 from http://www.cioinsight.com/article2/0,1540,1905930,00.asp

Layton, J., Brain, M., & Tyson, J. (2005). *How cell phones work.* Retrieved December 2005 from http://electronics.howstuffworks.com/cell-phone.htm

Hendler, J., Berners-Lee, T., & Miller, E. (2002). Integrating applications on the Semantic Web. *Institute of Electrical Engineers of Japan, 122*(10), 676-680.

Kowalczyk, R., Braun, P., Mueller, I., Rossak, W., Franczyk, B., & Speck, A. (2003). Deploying mobile and intelligent agents in interconnected e-marketplaces. *Journal of Integrated Design and Process Science, 7*(3), 109-123.

Lai, J., Mitchell, S., Viveros, M., Wood, D., & Lee, K. M. (2002). Ubiquitous access to unified messaging: A study of usability and the use of pervasive computing. *International Journal of Human-Computer Interaction, 14*(3/4), 385-404.

Lino, N. Q., Tate, A., Siebra, C., & Chen-Burger, Y.-H. (2003, August 11). Delivering intelligent planning information to mobile devices users in collaborative environments. *Proceedings of the 18th Joint Conference on Artificial Intelligence,* Acapulco, Mexico. Retrieved December 2005 from http://www.dimi.uniud.it/workshop/ai2ia/cameraready/queiroz.pdf

Mattern, F. (2000). *State of the art and future trends in distributed systems and ubiquitous computing.* Retrieved December 2005 from http://www.vs.inf.ethz.ch/publ/papers/DisSysUbiCompReport.pdf

Rodríguez, M., Favela, J., & Muñoz, M. A. (2003, August 11). Providing opportunistic access to information sources and services for mobile users. *Proceedings of the 18th Joint Conference on Artificial Intelligence,* Acapulco, Mexico. Retrieved December 2005 from http://www.dimi.uniud.it/workshop/ai2ia/cameraready/rodriguez.pdf

van Eijk, R., Hamers, J., Klos, T., & Bargh, M. S. (2002). *Agent technology for designing personalized mobile service brokerage.* Retrieved November 2005 from http://www.recursionsw.com/Mobile_Agent_Papers/Gigamobile.pdf

Wong, B. A., & Starner, T. E. (2001). *Conversational speech recognition for creating intelligent agents on wearables.* Retrieved November 2005 from http://www.cc.gatech.edu/fac/Thad.Starner/p/030_20_DPP/conversational-speech-recognition.pdf

Yeo, J., & Huang, W. (2003). Mobile e-commerce outlook. *International Journal of Information Technology & Decision Making, 2*(2), 313-332.

KEY TERMS

Bandwidth: The amount of data that can be transmitted in a fixed amount of time. For digital devices, the bandwidth is usually expressed in bits per second or bytes per second. For analog devices, the bandwidth is expressed in cycles per second, or Hertz (Hz).

Code-Division Multiple Access (CDMA): A digital cellular technology that uses spread-spectrum techniques. Unlike competing systems, such as GSM, CDMA does not assign a specific frequency to each user; instead, every channel uses the full available spectrum.

Distributed Network: A network structure in which the network resources, such as switching equipment and processors, are distributed throughout the geographical area being served.

Global System for Mobile Communications (GSM): One of the leading digital cellular systems. GSM uses narrowband Time Division Multiple Access technology, which allows eight simultaneous calls on the same radio frequency.

Neural Network: An interconnected collection of simple processing elements, units, or nodes whose functionality is loosely based on the animal brain. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns. Neural nets are used in bioinformatics to map data and make predictions.

3G: Third generation of mobile communications technology. 3G provides increased bandwidth up to 384 Kbps when a device is stationary or moving at pedestrian speed, 128 Kbps in a car, and 2 Mbps in fixed applications. 3G works over wireless air interfaces such as GSM or CDMA.

2G: Second-generation wireless service, also known as personal communications services. Refers to digital voice cell phone systems deployed in the 1990s. Delivers both voice and data transmissions using switched technology where each call requires its own cell channel.

Wireless Application Protocol (WAP): A standard protocol for providing cellular phones, pagers, and other handheld devices with secure access to e-mail and text-based Web pages. WAP provides a complete environment for wireless applications that includes a wireless counterpart of TCP/IP and a framework for telephony integration such as call control and phonebook access.

The Wi-INET Model for Achieving M-Health Success

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Steve Goldberg
INET International Inc., Canada

INTRODUCTION

Medical science has made revolutionary changes in the past decades. Contemporaneously however, healthcare has made incremental changes at best. The growing discrepancy between the revolutionary changes in medicine and the minimal changes in healthcare processes is leading to inefficient and ineffective healthcare delivery—one, if not *the,* significant contributor to the exponentially increasing costs plaguing healthcare globally. Healthcare organizations can respond to these challenges by focusing on three key solution strategies: access, quality, and value. These three components are interconnected such that they continually impact on the other, and all are necessary to meet the key challenges facing healthcare organizations today.

The application of mobile commerce to healthcare—namely, m-health—appears to offer a way for healthcare delivery to revolutionize itself and simultaneously address the critical areas of access, quality, and value. Integral to such an approach is the need for a robust wireless model. We propose the Wi-INET (wireless Internet, intranet, extranet) model as the way to deliver m-health excellence.

BACKGROUND

Currently the healthcare industry in the United States as well as globally is contending with relentless pressures to lower costs while maintaining and increasing the quality of service in a challenging environment. It is useful to think of the major challenges facing today's healthcare organizations in terms of the categories of demographics, technology, and finance. Demographic challenges are reflected by longer life expectancy and an aging population; technology challenges include incorporating advances that keep people younger and healthier; and finance challenges are exacerbated by the escalating costs of treating everyone with the latest technologies. Healthcare organizations can respond to these challenges by focusing on three key solution strategies: (1) *access*—caring for anyone, anytime, anywhere; (2) *quality*—offering world-class care and establishing integrated information repositories; and (3) *value*—providing effective and efficient healthcare delivery. These three components are interconnected such that they continually impact on the other and all are necessary to meet the key challenges facing healthcare organizations today. In short then, the healthcare industry is finding itself in a state of turbulence and flux (National Coalition on Healthcare, 2004; Pallarito, 1996; European Institute of Medicine, 2003; WHO, 2000, 2004; Wickramasinghe & Silvers, 2003). Such an environment, is definitely well suited for a paradigm shift with respect to healthcare delivery (von Lubitz & Wickramasinghe, 2005). Many experts within the healthcare field area agree that m-health appears to offer solutions for healthcare delivery and management that serve to maximize the value proposition for healthcare. However, to date, little if anything has been written regarding how to achieve excellence in m-health, nor does there exist any useful model for framing m-health delivery.

MAIN THRUST: INTEGRATIVE MODEL FOR M-HEALTH

Successful m-health projects require a consideration of many components. Figure 1 provides an integrative model for all key factors that we have identified through our research that are necessary in order to achieve m-health excellence (Wickramasinghe et al., 2005; Goldberg et al., 2002a, 2002b, 2002c, 2002d, 2002e; Wickramasinghe & Goldberg, 2004). What makes this model unique and most beneficial is its focus on enabling and supporting all areas necessary for the actualization of information and communication technology initiatives in healthcare. By design, the model identifies the inputs necessary to bring an innovative chronic disease management solution to market. These solutions are developed and implemented through a physician-led mobile e-health project. This project is the heart of the model to bridge the needs and requirements of many different players into a final (output) deliverable, a "Wireless Healthcare Program." To accomplish this, the model is continually updated to identify, select, and prioritize the ICT project inputs that will:

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- Accelerate healthcare system enhancements and achieve rapid healthcare benefits. The model identifies the key healthcare system inputs with the four Ps: ***people*** that deliver healthcare, ***process*** to define the current healthcare delivery tasks, ***platform*** used in the healthcare technology infrastructure, and ***protection*** of patient data.
- Close the timing gaps between information research studies and its application in healthcare operational settings.
- Shorten the time cycle to fund an ICT project and receiving a return on the investment.

IT Architecture and Standard Mobile Environment

By adopting a mobile/wireless healthcare delivery solution, it is possible to achieve rapid healthcare delivery improvements, which impact both the costs and the quality of healthcare delivery. This is achieved by using an e-business acceleration project which provides hospitals a way to achieve desired results within a standardized mobile Internet (wireless) environment. Integral to such an accelerated project is the ability to build on the existing infrastructure of the hospital. This then leads to what we call the three-tier Web-based architecture (see Figure 2).

In such an environment, Tier-1 is essentially the presentation layer; which contains the Web browser, but no patient data is stored within this layer, thereby ensuring compliance with international security standards/policies like HIPAA. Tier-2, shown as the HTTP Server, provides the business logic including but not limited to lab, radiology, and clinical transcription applications; messaging of HL7, XML, DICOM, and other data protocols; and interface engines to a hospital information system (HIS), lab information systems (LIS), radiology information systems (RIS), as well as external messaging systems such as Smart Systems for Health (an Ontario healthcare IT infrastructure project). This latter messaging feature may also be included in the third tier, which consists of the back-end database servers like Oracle, MySQL, or Sybase.

Mapping Case Study to Model

During the past six years, INET has used an e-business acceleration project to increase information and communication technology (ICT) project successes (Goldberg et al., 2002a, 2002b, 2002c, 2002d, 2002e). Today INET is repurposing the e-business acceleration project into a mobile e-health project to apply, enhance, and validate the mobile e-health project delivery model. Such a model provides a robust structure, and in turn serves to ensure excellence in the m-health initiative. INETs data provides the perfect opportunity to examine the components of our model (Figure 1), as it is both rich and longitudinal in nature. In mapping the data and specific business case, we have drawn upon many well-recognized

Figure 1. Wi-INET business model

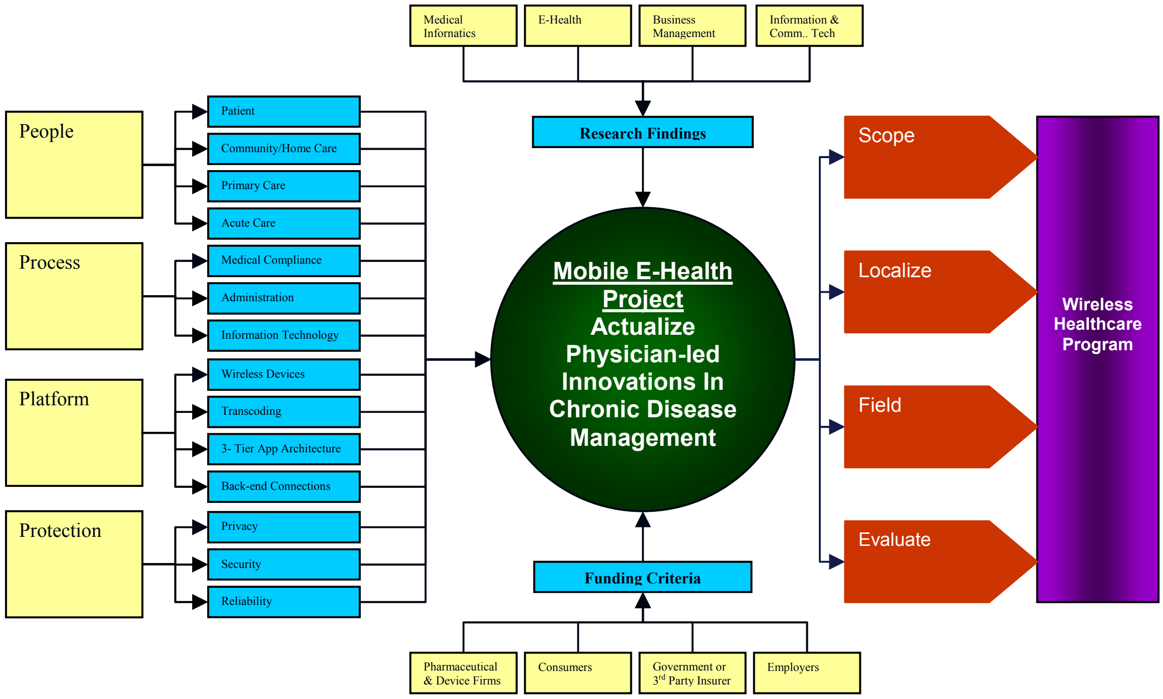
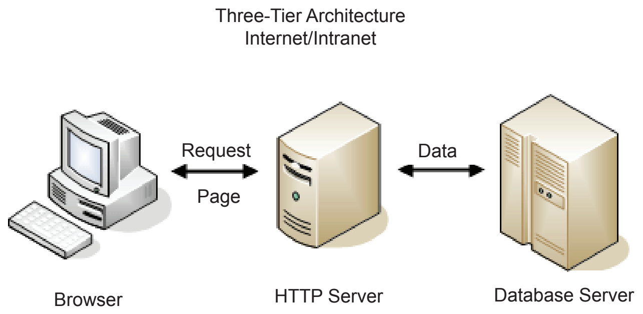

Medical Informatics
E-Health
Business Management
Information & Comm.. Tech
Research Findings
People
Patient
Community/Home Care
Primary Care
Acute Care
Process
Medical Compliance
Administration
Information Technology
Platform
Wireless Devices
Transcoding
3- Tier App Architecture
Back-end Connections
Protection
Privacy
Security
Reliability
Mobile E-Health Project
Actualize Physician-led Innovations In Chronic Disease Management
Scope
Localize
Field
Evaluate
Wireless Healthcare Program
Funding Criteria
Pharmaceutical & Device Firms
Consumers
Government or 3rd Party Insurer
Employers

Figure 2. Three-tier Web-based architecture



qualitative techniques, including conducting both structured and unstructured interviews, in-depth archival analysis, and numerous site visits.

Goldberg et al. (2002a, 2002b, 2002c, 2002d, 2002e) capture and substantiate the findings discussed, while Kavale (1996), Boyatzis (1998), and Einhardt (1989) detail the importance and richness of the methodologies we have adopted in presenting the following findings. Key criteria were established from the Standish Group International (1994) and the Institute of Medicine in America (2001).

The INET Mobile E-Health project objectives include:

1. Accelerate consensus building with an e-health solution that is focused on a disease state and driven by the medical model, with the primary objective to streamline communications and information exchange between patients, and providers of community/home care, primary care, and acute care.
2. Acquire commercial funding early with a compelling business case. For instance, enhancing therapeutic compliance can improve patient quality of life with significant healthcare cost savings. It is well documented that in diabetes, this will have immediate and high impact-benefits for healthcare consumers, pharmaceutical firms, governments, insurers, and employers.
3. Avoid risk by reengineering large-scale healthcare delivery processes in small manageable pieces. Today, organizations can harness a rigorous method to incrementally enhance a process one step at time. This is a way to achieve quick wins early and frequently.
4. Rapid development of simple-to-use, low-cost, and private/secure information and communication technology solutions. Achieve these benefits through a wireless application service provider (ASP). In addition to rapid development, a wireless ASP can easily connect and bring together many independent healthcare information systems and technology projects.

To actualize the mobile e-health project, INET is looking to the Wi-INET model as a framework. For INET, this will support an INET Mobile E-Health Project Management Office (PMO) to manage the costs and quality, deliver many small projects, and replicate projects for local and international distribution. As a first case scenario for the model, INET is proposing an INET Wireless Diabetes Program, with leadership from a family physician. The INET PMO is provisioning a project manager to support this physician-led project to meet both research and commercial sponsors' interests and objectives in diabetes. A detailed description of the key attributes of the INET Wireless Diabetes Program includes:

Problem Statement

There are many communication and information exchange bottlenecks between patients and their family physicians that prevent the effective treatment of diabetes. As background, a fundamental problem today is the ability to have a private and secure way to manage, search, and retrieve information at the point-of-care. In diabetes, physicians cannot quickly and easily respond to patients with high glucose levels. They need to wait for people to come to the office, respond to phone calls, reply using traditional mail delivery, or never receive the patient information.

Solution Mandate

Implement a diabetes monitoring program to enhance therapeutic compliance, such as release a program to enhance the usage of oral hypoglycemic agents (drugs) and/or the usage of blood sugar monitoring devices.

As background, everyone wins when enhancing patients' ability to follow instructions in taking prescribed medication. The patient's health, safety, and quality of life improve with significant healthcare cost savings. However, it is well documented that many patients do not stay on treatments prescribed by physicians.[1] This is where wireless technology may have the greatest impact to enhance compliance.

One solution may be as simple as using a cell phone and installing a secure wireless application for patients to

Table 1. INET wireless diabetes program results

INET Wireless Diabetes Program			
Patient	Change in HAIC Levels		% reduction in HAIC
	Pre-Pilot	Post-Pilot	
1	0.082	0.069	-16%
2	0.090	0.071	-21%
3	0.108	0.050	-54%
4	0.113	0.084	-26%

monitor glucose levels, and provisioning a physician to use a PDA (connected to a wireless network) to confidentially access, evaluate, and act on the patient's data.

Business Case

In Ontario the cost savings may represent almost $1 billion over three years. INET uses a simple calculation to determine the $1 billion savings. This can be found at www.inet-international.com; please select the INET mobile e-health project section to review the calculations.

The business case can be backed with additional data on how the cost of prevention (drugs) is far less than the cost savings associated with reducing the risk of complications associated with diabetes. For instance, the impact of a 1% decrease in A1C is significant. More data is available to support the business case for the prevention of type 2 diabetes such as lowering the incidence of End Stage Renal Disease (ESRD).

In summary, there is plenty of data today to quickly build consensus, fund, and implement a national and international wireless diabetes program to enhance patients' quality of life with significant healthcare cost reductions—that is, meet the objectives of access, quality, and value.

Systems Development Lifecycle Project Delivery

Use an INET mobile e-health project to scope, localize, field, and evaluate an INET wireless diabetes program led by a physician. Each project can easily and simply customize a program to quickly meet the unique needs of a rural and urban healthcare delivery setting, age, ethnicity, income, language, and culture. These are small manageable projects. Each project collects data on patient/healthcare provider relationships, wireless medical informatics, therapeutic compliance business case, and ICT usability to accelerate acceptance of a wireless diabetes program using wireless technology. The program may include cellular network and application usage, support, healthcare provider PDA, and consulting fees for a family physician and other healthcare providers. However, it is expected that the costs may not include items such as consumer cell phone, medication, or blood sugar monitoring devices/supplies. It is recommended that commercial and/or research sponsors pay for an INET project and help subsidize the user costs.

In June 2005, INET applied the Wi-INET model to pilot a wireless diabetes program with the objective to decrease diabetes-related complications with better control of glycemic levels, measured by HA1C.

The core component of the program is the relationship between family physicians and patients supported by a wireless diabetes management protocol.[2] This protocol describes how a patient can enter his or her glucose readings into a cell phone and transmit the results to his or her family physician. The protocol further details how the physician, in turn, is able to monitor any number of patients on his or her PDA, such as a Palm Treo or RIM Blackberry device. A physician, if required, can take immediate action with a message electronically sent to the patient's cell phone. The program was tested through a pilot project with four patients and led by Dr. Sheldon Silver, and was completed in July 2005. The pilot project lasted approximately three months. The preliminary results are significant as shown in Table 1.

In summary, INETs research data indicates that using the Wi-INET model will increase ICT project success in healthcare. To realize and test this, INET continues to map the player's from an INET wireless diabetes program (use case scenario) to the model. To show how this works, please review the mapping exercise below. The bold text in black is a project player and the color text in [] parenthesis relates to the sections of the model presented in Figure 1.

Physician Mobile E-Health Project Lead: ["Mobile E-Health Project" in Figure 1]. Physicians provide the linkage to the medical model to enhance disease management programs to enhance patient care and safety, improve research and education, increase healthcare quality, and reduce healthcare costs. For INET's use case scenario, the final outcome is a Wireless Diabetes Program ["Wireless Healthcare Program" in Figure 1]. And the mobile e-health projects are led by Dr. Sheldon Silver, MD, Staff Physician, Credit Valley Hospital.

Commercial Sponsor(s) ["Funding Criteria" in Figure 1]. The project delivers information and communication solutions for:

- Consumers wishing to improve their quality of life with an enhanced relationship with their healthcare provider's—that is, family physicians.
- Pharmaceutical firms looking to increase revenues with e-compliance programs.
- Government/insurers investigating ways to significantly reduce administration and healthcare costs, and shorten healthcare delivery time cycles (wait times.)

- Employers wanting to increase productivity and avoid absenteeism with a healthier workforce.

Research Sponsor(s) ["Research Findings" in Figure 1]: The project develops intellectual property for researchers in the fields of:

- Patient and Healthcare Provider Relationships
- Wireless Medical Informatics
- Therapeutic Compliance Business Case
- Wireless Information Technology Usability

An INET Mobile E-Health Project Delivery Team:

Healthcare Delivery Team ["People" in Figure 1]: For a wireless diabetes program the players may include:

- Healthcare Consumer: People with Diabetes
- Community Care: Nurse Specializing in Diabetes
- Primary Care: Family Physician
- Acute Care: Endocrinologist and Diabetes Education/Management Centers

Business Process Analyst ["Process" in Figure 1]

Privacy and Security Consultant ["Protection" in Figure 1]:

Programmer using a wireless ASP ["Platform" in Figure 1]:

- Wireless Network and Devices
- Device and Application Transcoding
- Application Service Provider
- Back-End Connection

In conclusion, INET is looking forward to further advancements in the mobile e-health project delivery model to:

- achieve rapid advancements in healthcare delivery,
- improve diabetes management,
- enhance therapeutic compliance, and
- realize significant healthcare care cost savings.[3]

INET is planning to continue its role as a source of use case scenarios for the model with the delivery of mobile e-health projects.

FUTURE TRENDS AND CRITICAL SUCCESS FACTORS

INET's experience and preliminary findings are pointing to one critical success factor—the commercialization of physicians' innovations in chronic disease management. INET is expecting significant growth in this healthcare trend, with wireless technology as the fundamental enabler. With antidotal comments from patients that participated in the INET pilot project are testimonials to this new era in healthcare delivery. A few of their comments (*Canadian Healthcare Technology,* 2005) include:

"All I had to do was pick up my phone and dial, punch in my ID, put in the number for my blood sugar level, and hit enter. That was it. Did it twice a day," said patient Jim Pott at the INET conference. "It gave me the feeling of constantly being looked after because, if I forgot to do it, Dr. Silver would be on my case, sending a text message to remind me." For patient Dave Rowan: "I'm in the computer industry, so using an input device like my Treo phone is something I am very used to. Even so, to have it confirmed that my information was received and have subsequent interaction with Dr. Silver was terrific. And I could do it from all over North America." For Joy Merritt, the trial has been life transforming. "Dr. Silver saved my life with this system. I felt totally out of control before. But now I feel I can control my life. And because I see my levels every day, and I know the doctor does too, that's motivated me to exercise more and try to bring them down."

The preceding has served to outline all the critical aspects that must be considered when trying to actualize a mobile health initiative. Clearly, mobile e-health or m-health projects are complex and require much planning and coordination within and between the web of healthcare players. Success is never guaranteed in any large initiative, however in order to realize the four major healthcare deliverables depicted in Figure 1 (enhance patient care and safety, improve research and education, increase healthcare quality, and reduce healthcare costs), it is vital that any m-health initiative focus on the key success factors of people process and technology. Specifically, the technology must be correct and functioning as desired. Further, it must integrate seamlessly with existing ICT infrastructure and enable the processes. The processes must be well defined and at all times ensure that they are of a high quality and error free.

The Institute of Medicine in America (2001) identified medical errors as the fourth leading cause of many deaths. In trying to prevent such errors, it has identified six key quality aims:

1. **Healthcare Should be Bafe:** Avoiding injuries to patients from the care that is intended to help them.
2. **Healthcare Should be Effective:** Providing services based on scientific knowledge to all who could benefit, and refraining from providing services to those who will not benefit (i.e., avoiding under use and overuse).
3. **Healthcare Should be Patient-Centered:** Providing care that is respectful of and responsive to individual

patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions.

4. **Healthcare Should be Timely:** Reducing waiting and sometimes harmful delays for both those receiving care and those who give care.
5. **Healthcare Should be Efficient:** Avoiding waste.
6. **Healthcare Should be Equitable:** Providing care that does not vary in quality based on personal characteristics.

Finally, and arguably the most critical key success factor, is the web of healthcare players. Any m-health project must consider the impact and role on each of these players, the interactions of such an initiative both within one group of the web of players as well as between groups of players. As discussed, and based on the findings from INETs longitudinal studies, it is critical to the ultimate success of these projects and the ability to in fact realize the healthcare deliverables that they are indeed physician led.

DISCUSSION AND CONCLUSION

Healthcare in the United States and globally is at the crossroads. It is facing numerous challenges in terms of demographics, technology, and finance. The healthcare industry is responding by trying to address the key areas of access, quality, and value. M-health, or mobile e-health, provides a tremendous opportunity for healthcare to make the necessary evolutionary steps in order to realize its goals and truly achieve its value proposition. What is important is to ensure m-health excellence. This requires not only detailed theoretically studies, but ultimately the need to turn theory into practice. By an in-depth analysis of the rich and longitudinal data of INET, we have developed the Wi-INET model to facilitate the achievement of m-health excellence. Systematic and detailed analysis and integration of all the key drivers and implication of healthcare delivery have enabled the development of the Wi-INET model. Moreover its structure facilitates rapid development and actualization of m-health solutions. To the best of our knowledge, it is the first such model, and while it is certainly not a panacea, it does help to set the stage and outline the key issues that must be addressed for a successful m-health initiative, and it enables healthcare to reap the benefits of wireless. We are confident that through the adoption of the Wi-INET model, healthcare delivery too can make revolutionary changes and we can all enjoy superior healthcare delivery.

REFERENCES

Blair, J. (2004). Assessing the value of the Internet in health improvement. *Nursing Times, 100,* 28-30.

Boyatzis, R. (1998). *Transforming qualitative information thematic analysis and code development.* Thousand Oaks, CA: Sage.

Canadian Healthcare Technology. (2005). Wireless reporting for diabetes patients offers up dramatic results. *Canadian Healthcare Technology,* (September), 4. Retrieved from http://www.inet-international.com/INET/Update/PressCoverage2005.htm

Eisenhardt, K. (1989). Building theories from case study research. *Academy of Management Review, 14,* 532-550.

European Institute of Medicine. (2003). *Health is wealth: Strategic vision for European healthcare at the beginning of the 21st century.* Salzburg, Austria: European Academy of Arts and Sciences.

Frost & Sullivan Country Industry Forecast. (2004). *European Union healthcare industry.* Retrieved May 11, 2004, from http://www.news-medical.net/print_article.asp?id=1405

Goldberg, S. et al. (2002a, January). *Building the evidence for a standardized mobile Internet (wireless) environment in Ontario, Canada.* Internal INET Documentation.

Goldberg, S. et al. (2002b). *HTA presentational selection and aggregation component summary.* Internal Documentation.

Goldberg, S. et al. (2002c). *Wireless POC device component summary.* Internal INET Documentation.

Goldberg, S. et al. (2002d). *HTA presentation rendering component summary.* Internal INET Documentation.

Goldberg, S. et al. (2002e). *HTA quality assurance component summary.* Internal INET Documentation.

INET Talk. (2004, October 14). Enhance therapeutic compliance using wireless technology. *Proceedings of the WNY Technology & Biomedical Informatics Forum,* Niagara Falls, NY.

Institute of Medicine in America. (2001). *Crossing the quality chasm: A new health system for the 21st century.* Committee on Quality of Healthcare in America, Institute of Medicine. Washington, DC: National Academy Press.

Kavale, S. (1996). *An introduction to qualitative research interviewing.* Thousand Oaks, CA: Sage.

Kulkarni, R, & Nathanson, L.A. (2005). *Medical informatics in medicine.* Retrieved from http://www.emedicine.com/emerg/topic879.htm

Kyprianou, M. (2005). The new European healthcare agenda. *Proceedings of the European Voice Conference: Healthcare: Is Europe Getting Better?* Retrieved from http://www.noticias.info/asp/aspcommunicados.asp?nid=45584

Lacroix, A. (1999). International concerted action on collaboration in telemedicine: G8sub-project 4. *Sted. Health Technol. Inform, 64,* 12-9.

Lee, M. Y., Albright, S. A., Alkasab, T., Damassa, D. A., Wang, P. J., & Eaton, E. K. (2003). Tufts Health Sciences Database: Lessons, issues, and opportunities. *Academic Medicine, 78,* 254-264.

National Center for Health Statistics. (2002). *Health expenditures 2002.* Retrieved from http://www.cdc.gov/nchs/fastats/hexpense.htm

National Coalition on Healthcare. (2004). *Building a better health: Specifications for reform.* Washington, DC: National Coalition on Healthcare.

Organisation for Economic Cooperation and Development. (2004). *OECD health data 2004.* Retrieved from www.oecd.org/health/healthdata

Pallarito, K. (1996). Virtual healthcare. *Modern Healthcare,* (March), 42-44.

Plunkett's. (2005). *Plunkett's health care industry almanac.* Houston: Plunkett Research.

Russo, H.E. (2000). The Internet: Building knowledge & offering integrated solutions to health care. *Caring 19,* 18-20, 22-24, 28-31.

Standish Group International. (1994). *The CHAOS report.* Retrieved from http://www.standishgroup.com/sample_research/chaos_1994_1.php

von Lubitz, D., & Wickramasinghe, N. (2005). Healthcare and technology: The doctrine of network-centric healthcare. *Health Affairs.*

WHO. (2000). *Health systems: Improving performance* (pp. 1-215). Geneva: World Health Organization.

WHO. (2004). *Changing history* (pp. 1-167). Geneva: World Health Organization.

Wickramasinghe, N., & Goldberg, S. (2004). How M=EC2 in healthcare. *International Journal of Mobile Communications, 2*(2), 140-156.

Wickramasinghe, N., & Mills, G. (2001). MARS: The electronic medical record system, the core of the Kaiser galaxy. *International Journal of Healthcare Technology Management, 3*(5/6), 406-423.

Wickramasinghe, N., & Silvers, J.B. (2003). IS/IT: The prescription to enable medical group practices to manage managed care. *Health Care Management Science, 6,* 75-86.

Wickramasinghe, N. et al. (2005). Assessing e-health. In T. Spil & R. Schuring (Eds.), *E-health systems diffusion and use: The innovation, the user and the user IT model.* Hershey, PA: Idea Group Publishing.

KEY TERMS

Healthcare Challenge: One of the challenges that can be thought of in terms of demographic issues such as the aging population, technology issues such as the need to embrace technologies into healthcare, and finance issues such as escalating healthcare costs.

Key Healthcare System Input: Includes people, process, platform, and protection.

M-Health: The application of wireless technology to healthcare delivery.

Mobile Healthcare Business Model: A model to identify all the key drivers and actors for a mobile healthcare initiative.

Mobile Healthcare Delivery Model: A model that outlines the process and procedures to realize the outlined business model.

Superior Healthcare Delivery: A patient-centric healthcare delivery system that tries to maximize access, quality, and value.

Three-Tier Web-Based Architecture: ICT backbone of the wireless m-health initiative.

Wireless Healthcare Program: The incorporation of wireless technology into a specific healthcare program, for example, the use of wireless technology in a diabetes program.

ENDNOTES

[1] Fourteen percent to 21% of patients never fill their original prescription, and 10% to 50% of patients ignore or otherwise compromise their medication instructions (http://www.managedhealthcareexecutive.com/mhe/article/articleDetail.jsp?id=105388).

[2] Wireless Diabetes Management Protocol ©Dr. Sheldon Silver, MD, 2005.

[3] In Ontario, this may save 1 $billion over three years (INET Talk, 2004).

Wireless Access Control System Using Bluetooth

Juliano Rodrigues Fernandes de Oliveira
Federal University of Campina Grande, Brazil

Rodrigo Nóbrega Rocha Xavier
Federal University of Campina Grande, Brazil

Yuri de Carvalho Gomes
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

Security is one of the world's main challenges. Research and industrial applications related to security include several areas such as personal security, organizational security, and computer security, among others. This article is concerned with secure environments, which is related to the control of people entering an environment, building, rooms, laboratories, and so forth. In this context, access control systems are the main security mechanisms to control the access of authorized people to environments.

Nowadays, locks and keys are not enough to keep an environment secure against unwanted or uncontrolled visitors. To have access, mechanical security systems are widely used, however, such systems—purely mechanical—can be easily defrauded. To construct high-security access systems, the embedded electronics have associated to the mechanical security, with the objective of increasing the level of reliability of such systems. Besides, with the increasing use of mobile devices, users are more and more interested in mobile solutions to support several activities, including security-related ones.

This article presents an access control system that uses Bluetooth technology (Ericsson Bluetooth, 2006) to allow control of the entrance to environments. By using the proposed system, a person with a smart phone can use it to get access to environments, such as buildings, labs, rooms, and so forth.

The remainder of this article is organized as follows. First we present the architectural components of the proposed system and detail their functioning. We then discuss future trends and offer concluding remarks.

BACKGROUND

Bluetooth

The Bluetooth specification was developed by Ericsson (now Sony Ericsson) and later formalized by the Bluetooth Special Interest Group (SIG). The SIG was formally announced on May 20, 1999, and originally founded by Ericsson, IBM, Intel, Nokia, and Toshiba.

Bluetooth is an industrial standard for wireless personal area networks (PANs), also known as IEEE 802.15.1 (Bluetooth SIG, 2004). It provides a secure, low-cost way to connect and exchange information between devices, such as personal digital assistants (PDAs), mobile phones, laptops, PCs, printers, and digital cameras, in a globally available short-range radio frequency. This technology eliminates cables and wires between devices, facilitates both data and voice communication, and enables ad-hoc networks between multiple Bluetooth devices (Cardei, 2002).

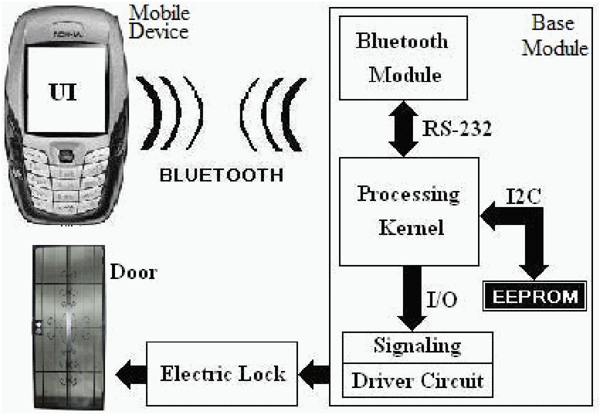
Bluetooth is a radio standard primarily designed for low power consumption, with a short range (power class dependent: 1 meter, 10 meters, 100 meters) and with a low-cost transceiver microchip in each device. It lets these devices communicate with each other when they come in range, even if they are not in the same room, as long as they are within up to 100 meters of each other, depending on the power class of the product (Kardach, 1998).

Microcontrollers

A microcontroller (MCU) is a computer-on-a-chip used to control electronic devices. It is a microprocessor emphasiz-

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Wireless access control system architecture



ing self-sufficiency and cost-effectiveness, in contrast to a general-purpose microprocessor used in a PC. It can be defined as a single integrated circuit with a central processing unit, usually small and simple; input/output interfaces, such as serial ports; peripherals, such as timers and watchdog circuits; RAM for data storage; ROM for program storage; and a clock generator, often an oscillator for a quartz timing crystal, resonator, or RC circuit (Stewart, 1993).

In addition to the key features, most microcontrollers today take further advantage of not needing external pins for memory buses. They can afford to use the Harvard architecture: separate memory buses for instructions and data, allowing multiple access to occur concurrently (Cady, 1997).

A typical microcontroller contains all memory and interfaces needed for a simple application, whereas a general purpose microprocessor requires additional chips to provide these functions. Microcontrollers also usually have a variety of input/output interfaces. Serial I/O (UART) is very common, and many include analog-to-digital converters, timers, or specialized serial communications interfaces like I^{2}C, serial peripheral interface (SPI), and controller area network (CAN).

A microcontroller is also a programmable device that can be destined for several purposes. The firmware recorded in its memory is responsible for the characteristic of its application. Microcontrollers are versatile tools and with low cost for embedded systems design.

Originally, microcontrollers were only programmed in assembly language, or later in C code. Recent microcontrollers, integrated with on-chip debug circuitry accessed by in-circuit emulator via JTAG, enable a programmer to debug the software of an embedded system with a debugger (Cady, 1997).

Microcontrollers trade speed and flexibility against ease of equipment design and low cost. This integration drastically reduces the number of chips and the amount of wiring and space that would be needed to produce equivalent systems using separate chips. Manufacturers and designers have to balance the need to minimize the chip size against additional functionality.

SYSTEM ARCHITECTURE

The access control system architecture depicted in Figure 1 consists of two modules: mobile and base. A smart phone contains the software responsible for beginning the authentication process, acting as mobile module. The base module is responsible to receive a valid authentication code and to allow the access to the environment by unlocking an electric lock embedded in the environment entrance door.

The base module is composed of a Bluetooth module (Wintec BT Module, 2005); a processing kernel, represented by a microcontroller (Microchip PIC18FXX2, 2002); an external data storage unit, represented by an EEPROM memory (Microchip 24LC256, 2002); and an electric lock interface, represented by a driver circuit to unlock the electric lock.

In general, the user authentication process consists of sending the user authentication key from the application running in a mobile device to the Bluetooth module, through Bluetooth connection. The Bluetooth module sends such information to the processing kernel, which performs the authentication through comparison of the user key sent with that stored in the external data storage unit. Next, the processing kernel sends the search authentication result to the mobile device and to the electric lock interface. If the user key is valid, the electric lock interface unlocks the environment entrance door. Each architectural component is detailed in what follows.

Mobile Module and Bluetooth Module

Mobile module is the application embedded in a mobile device that performs the communication with the Bluetooth module. It has been developed in J2ME language (J2ME, 2006). Such an application is based on the Bluelet open source software (Bluelet, 2006). The entire connection negotiation process has been implemented using protocols of the Bluetooth protocol stack to perform connections via Serial Port Profile (Wintec Bluetooth, 2004).

The basic process to connection negotiation consists of three steps:

1. Search for the Bluetooth module (discovery function), through the name "Wintec Serial Port" or the Bluetooth module address.
2. Authentication, or pairing, using the code sent by the mobile device to the Bluetooth module (bond function).

3. Connection establishment (connect function) based on serial port profile. A wireless communication is emulated by a connection via serial port with UART protocol (Wintec Bluetooth, 2004).

Regarding the application functioning, first of all an initial connection attempt is performed to connect to the Bluetooth Module directly, through the Bluetooth module address. Such an address is discovered by localization of Bluetooth devices and stored in the device. If the Bluetooth module is inaccessible—distant or turned off—or if it is not found in the direct connection attempt, the software automatically performs two more attempts, notifying the user visually. If no attempt works, then the software will be finished.

After the Bluetooth connection establishment, the access control system awaits the sending of the user key. In this case, a string containing the user authentication key is transmitted from the smart phone to the Bluetooth module. This key needs to be registered in the application by the user and then stored in the mobile device. If the operation is successful, the electric lock is unlocked and the user is informed visually that the door has been opened. Afterwards, the connection is closed and the application is finished.

Processing Kernel

During the connection establishment process between the smart phone and the Bluetooth module, messages (in ASCII format) are sent from the Bluetooth module to the microcontroller host, which represents the Processing Kernel. The firmware contained in the microcontroller host monitors these messages awaiting the final message of connection closing. Meanwhile, it continues awaiting the user authentication key that must be sent by the mobile device. When it receives such a key, it analyzes it, and if the key is registered in the system database, the host will unlock the electric lock. After the data exchange between the mobile device and the Bluetooth module, the microcontroller sends to the Bluetooth module the escape sequence to close the Bluetooth connection.

External Data Storage Unit

The external memory EEPROM is used as a persistent data storage device in this wireless access control system. It is very important due to the reduced capacity of internal memory available in the microcontroller.

The user information is stored in the external memory as a *login*/*password* table. A pointer to free memory positions is used to indicate which memory spaces are available to register new users.

The recording operation of user authentication keys begins when the processing kernel receives a *write command.* The user information, *login* and *password,* contained in this command are then stored in the external memory unit.

The search operation for user authentication keys begins when the processing kernel receives an *access command.* The user information, *login* and *password,* contained in this command are acquired and stored in vectors, for search and comparison with stored data in the memory unit. If user information is valid, the operation will be successful. Otherwise, the operation has failed and the electric lock will not be unlocked.

Electric Lock Interface

The electric lock interface is represented by a driver circuit of the electric lock and by a panel composed by LEDs that indicates the current operations during the authentication process. It is responsible for informing of the status of the driver circuit. There are four LEDs. One of them indicates that the circuit is energized and active. Another indicates that a recording operation is currently being performed. Yet another LED indicates that the search operation for user key has been performed successfully and the electric lock was unlocked, allowing access to the environment. The last LED indicates that the search operation has failed (invalid code) and the electric lock has not been unlocked, not allowing the user access to the environment.

FUTURE TRENDS

In the context of the proposed system, in order to improve the reliability of the data exchange between the remote device and the access control system, revisions in the firmware contained in the processing kernel are necessary. Some suggestions to increase the reliability are: addition of an administrator user, development of new functions for this administrator to control and supervise the system, and better sub-routines to search for registered users.

Regarding future efforts in mobile-related access control technologies, the main trends are concerned with pervasive environments. For example, current access control technology works by keeping the entrance door closed and opening it for authorized persons. But there is another way: the door could be left open and only closes when an unauthorized person tries to enter. In this case, the access control system must be monitoring the environment, and when an unauthorized person comes close, the door is blocked. This example has a pervasive characteristic in which the system is "invisible" for the user. Several access control researches are moving in that direction, to conceive environments that manage and control their security without needing a direct user intervention.

CONCLUSION

Security is a growing need throughout the world, and lack of security can result in great damage. Many solutions are available for all levels of access control—from highly restricted areas such as laboratories or computer rooms to less restricted areas such as storage rooms.

Access control solutions include electronic keys, magnetic stripe cards, proximity cards, and smart cards or biometric devices, including hand and fingerprint readers. More sophisticated access control capabilities, such as auto-unlock/auto-lock functionalities, allow programming an electronic locking system to lock and unlock any door at any time.

With the increasing personal use of mobile devices and growing industry investments in this area, it is a trend to use mobility capabilities to support user activities, mainly security-related ones. The proposed access control system using Bluetooth is a good example of how to join mobility and reliability to support user activities in a practical and secure way.

REFERENCES

Bluelet. (2006). *Bluetooth GUI component.* Retrieved from http://benhui.net/

Bluetooth SIG. (2004). *Bluetooth Special Interest Group launches Bluetooth Core Specification version 2.0 + Enhanced Data Rate.*

Cady, F. M. (1997). *Microcontrollers and microcomputers.* Oxford: Oxford University Press.

Cardei, M. (2002). Overview over the Bluetooth technology. *Proceedings of the Wireless Networking Seminar,* University of Minnesota.

Ericsson Bluetooth. (2006). *Ericsson Bluetooth.* Retrieved from http://www.ericsson.com.br/bluetooth/index.asp

J2ME. (2006). *Java 2 Micro Edition.* Retrieved from http://java.sun.com/j2me/index.jsp

Kardach, J. (1998). *Bluetooth architecture overview.* Intel.

Microchip 24LC256. (2002). *Microchip 24LC256 256k I2C CMOS Serial EEPROM datasheet.* Microchip Technology.

Microchip PIC18FXX2. (2002). *Microchip PIC18FXX2 datasheet.* Microchip Technology.

Stewart, J.W. (1993). *The microcontroller.* Englewood Cliffs, NJ: Regents/Prentice-Hall.

Wintec Bluetooth. (2004). *Wintec Bluetooth SPP command interface quick start guide.* Wintec Industries.

Wintec BT Module. (2005). *WBTV42-D-XXX Bluetooth module rev 0.8 guide.* Wintec Industries.

KEY TERMS

American Standard Code for Information Interchange (ASCII): A standard for coding text files. Every character has an associated number, and any text can be represented by a sequence of numbers.

Electrically Erasable Programmable Read-Only Memory (EEPROM): A non-volatile storage chip used in computers and other devices (such as USB flash drives, in its flash memory version); also called E2PROM.

I^{2}C (Inter-IC) Bus: A bi-directional two-wire serial bus that provides a communication link between integrated circuits (ICs).

J2ME: Collection of Java APIs for the development of software for resource-constrained devices such as PDAs, cell phones, and other consumer appliances.

Serial Port Profile (SPP): Defines how to configure and connect the virtual serial port between two wireless devices supporting Bluetooth technology.

Smart Phone: Any electronic handheld device that integrates the functionality of a mobile phone, personal digital assistant (PDA), or other information appliance. This is often achieved by adding telephone functions to an existing PDA or putting "smart" capabilities, such as PDA functions, into a mobile phone. A key feature of a smart phone is that additional applications can be installed on the device. The applications can be developed by the manufacturer of the handheld device, by the operator, or by any other third-party software developer.

Symbian OS: An operating system designed for mobile devices, with associated libraries, user interface frameworks, and reference implementations of common tools, produced by Symbian Ltd.

UART: Universal asynchronous receiver/transmitter protocol.

Wireless Client Server Application Model Using Limited Key Generation Technique

Rohit Singh
Monash University, Australia

Dhilak Damodaran
Monash University, Australia

Phu Dung Le
Monash University, Australia

INTRODUCTION

The introduction of personal digital assistants (PDAs) and laptops has brought in mobility for carrying out computing jobs (Pasquale, Hung, Newhouse, Steinberg, & Ramabhadran, 2002; Varshney & Vetter, 2000). Wireless networks in working areas cater the need for installation flexibility, reduced cost-of-ownership, mobility and scalability. Mobile computing is distinguished from classical, fixed-connection computing due to the mobility of nomadic users and their devices (Jing, Helal, & Elmagarmid, 1999) and mobile environment is defined as low bandwidth and high latency networks with devices supporting limited input methods and containing low power processors, small display size and short battery life (Buszko, Lee, & Helal, 2001). Mobile working environments attract the users, employees and students, but presents hardships to the security programmer because the settings and the environment of mobile devices vary significantly from that of wired devices.

The ease of network access should be combined with reliable security services. Wireless networks are exposed to the security compromises because it provides hacker easy access to transport media. An attacker can sniff the packets by setting up equipment monitoring at 2.4GHz frequencies and capable of interpreting the packets of 802.11 standard (Borisov, Goldberg, & Wagner, 2001; IEEE Standard 802.11b, 1999; Josang & Sanderud, 2003). This unauthorized access to the network is a matter of great concern for any network security administrator. Cryptography, that is encryption of the data, is one of the best ways to provide security to wireless communication. Even the strongest of the encryption procedures can become vulnerable to possible attacks when the keys of the parties get compromised. Thus the lack of security is predominantly due to poor management of keys rather than the weakness in the encryption algorithm (Josang & Sanderud, 2003).

In this article we implement a client server model using limited-used key generation scheme (Kungpisdan, Le, & Srinivasan, 2004) to generate a set of session keys that are never transmitted, which means that there is no chance for the attacker to sniff the packets and retrieve keys while they are being transmitted. These session keys are used for encrypting and hashing the data to be transmitted from mobile client device to the servers in wired network and vice versa. The updating of the session keys used in this technique does not rely on any long-term shared key, instead the process is based upon the last session key used. This technique of elevating the frequency of the key update to the next possible level makes the system much more secure than the other present techniques. In addition to providing better security, this technique also enhances the performance of a limited resource device by avoiding the repeated generation of keys on it.

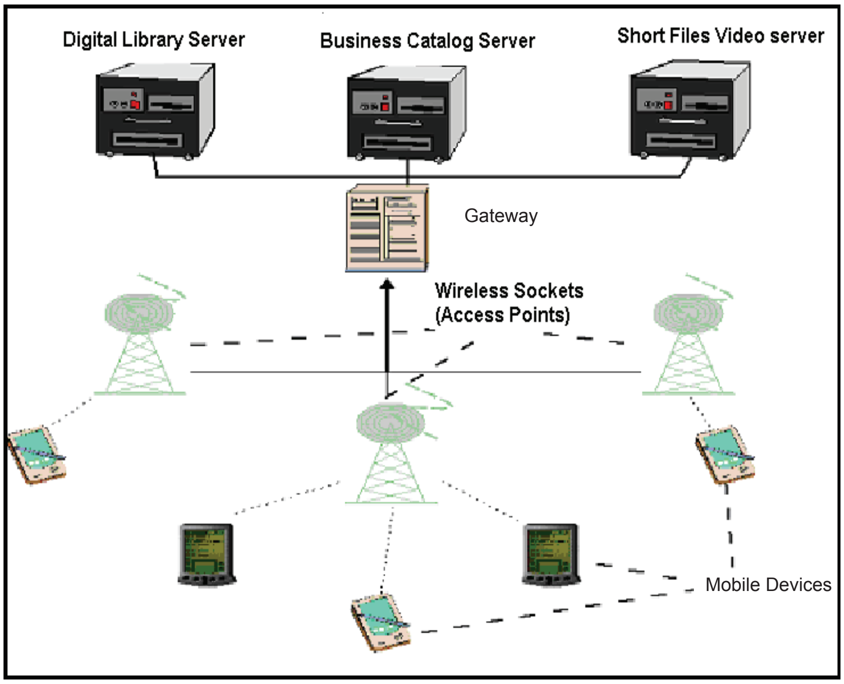
The rest of the article is organized as follows. The second section gives an overview of the communication between mobile devices and the various servers running in the wired network. The third section describes the proposed technique. The fourth section discusses the technique in applied state. Next, the fifth section discusses about other key management and security issues of the technique. The last section concludes the article.

CLIENT SERVER MODEL

The client server model in a wireless environment is depicted in Figure 1. Mobile clients in wireless networks first connect to an access point, which is connected with wired media to the main network infrastructure. After a successful connection with the access point, mobile clients can start communicating with the gateway and servers that are present within the main network. The model described in this article proposes to set up a gateway program which authenticates every mobile client before it connects to any server inside the main network. The authentication is carried on the basis of the technique discussed in the fourth section. The gateway program can be executed at the gateway server of organization.

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Remote access model



The course of action that is carried out in the proposed protocol is as follows:

- **C → G:** Server Connection (Request), Command Execution (Request)
- **G → S:** Command-Execution (Request)
- **S → G:** Command-Execution (Response)
- **G → C:** Server Connection (Response), Command-Execution (Response)

PROPOSED TECHNIQUE

Notations

- $\{C, G, S\}$: the set of clients, gateway and server, respectively.
- $\{K_A, K_A^{-1}\}$: the set of public/ private key for party A.
- $\{ID_C, ID_S, ID_G\}$: the set of identities of client, the server and the gateway respectively.
- $ID_G Req$: request for ID_G
- *CI*: Command execution information.
- *CRes*: Command response
- $\{M\}_X$: the message M symmetrically encrypted with the shared key *X*.
- $\{M\}_{Kx}$: the message M encrypted with the public key of the party *X*.
- $\{M\}_{Kx}^{-1}$: the message M signed with the private key of the party *X*.
- $h(M)$: the one-way hash function (FIPS, 1995) of the message *M*.
- $MAC(M, K)$: the message authentication code (*MAC*) of the message *M* with the key *K*.

Initial Assumption and Settings

Initial assumptions and settings for the proposed technique are as follows:

1. Client is considered to be a person with the wireless device and has access to the organization's network.
2. Client's employment record (CER), containing employee code and other information about the client, is the long-time shared secret between the server and Client. CER is assumed to be a key that never expires.
3. The distributed key DK is another shared key between client and server. This key is distributed by performing authenticated key exchange (AKE) (Boyd & Park, 1998; Horn & Preneel, 1998; Kungpisdan, Le, & Srinivasan, 2003; Toh, Kungpisdan, & Le, 2004; Wong & Chan, 2001; Zhu, Wong, Chan, & Ye, 2002) protocol between client and server.

C → S: $\{ID_C, DK, n\}_k$

S → C: $\{n\}_k$

This key is then further used to generate session key Y_i using the technique explained in the third section.

4. There is another distributed key X that is a shared secret between client and gateway. This key is transferred to the gateway by performing AKE protocol.

C → G: $\{ID_C, X, n\}_k$

G → C: $\{n\}_k$

This key X is then further used to generate session key X_i using the technique explained in 3.3.2.

5. h (M, K) stands for the key hashed function for the message M and key K. For higher security reason HMAC is preferred.

Key Generation Technique

Generating Y_i: Figure 2 shows the procedure to generate Y_i. The steps to generate keys on the communicating terminals are as follows:

1. Client generates the distributed key and transfers the same to the server through a secure channel. Client and server generate a set of preference keys K_i, where i=1,2,3,….*m*:

$K_1 = h\ (DK, CER),\ K_2 = h\ (DK, K_1)\ldots\ldots,$
$K_m = h\ (DK, K_{m-1})$

CER is then removed from the system. This type of recursive hashing is represented by *DK*CER* in remaining part of paper.

2. Client generates a random number *r*, which is sent to the server at the start of every session.

C → S: *r*

3. Client selects two preference keys K_{mid1} and K_{mid2}. K_{mid1} is the middle key among $\{K_1, K_2, K_3\ldots\ldots, Kw\}$, where $w = r \bmod m$ and K_{mid2} is the middle key among $\{K_1, K_2\ldots. K_{mid1}\}$. Using these keys session initialization key (*SIK*) is calculated.

SIK= h (K_{mid1}, K_{mid2})

After generating *SIK*, K_{mid1} and K_{mid2} are removed from the terminals. As the value of *r* was transmitted over to the server, the same *SIK* is generated at the server's terminal as well.

4. Server and client generate a set of session keys Y_i, where *i*=1, 2…, *n*.

$Y_1 = h\ (SIK, DK),\ Y_2 = h\ (SIK, Y_1)\ldots\ Y_n = h\ (SIK, Y_{n-1})$

These session keys are then used by server and client to encrypt, decrypt and hash the data passed between them.

Figure 2. Generation of Y_i

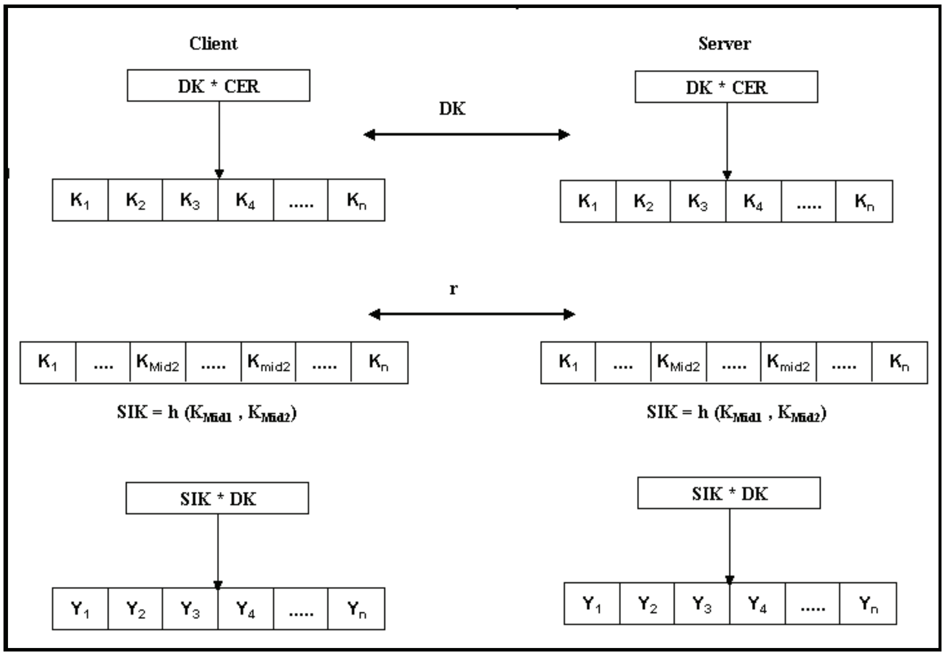
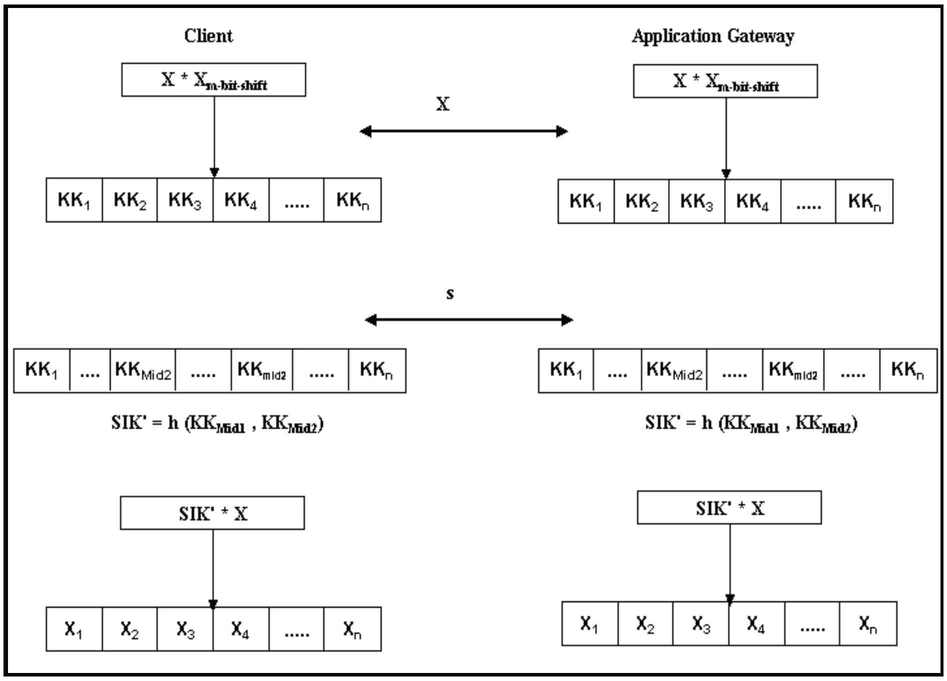


Figure 3. Generation of X_i


Client
Application Gateway
$X * X_{m\text{-}bit\text{-}shift}$
$X * X_{m\text{-}bit\text{-}shift}$
X
KK_1 KK_2 KK_3 KK_4 ..... KK_n
KK_1 KK_2 KK_3 KK_4 ..... KK_n
s
KK_1 .... KK_{Mid2} ..... KK_{mid2} ..... KK_n
KK_1 .... KK_{Mid2} ..... KK_{mid2} ..... KK_n
$SIK' = h(KK_{Mid1}, KK_{Mid2})$
$SIK' = h(KK_{Mid1}, KK_{Mid2})$
$SIK' * X$
$SIK' * X$
X_1 X_2 X_3 X_4 ..... X_n
X_1 X_2 X_3 X_4 ..... X_n

The application of the session keys in proposed technique is discussed in the fourth section.

Generating X_i: Figure 3 shows the procedure to generate X_i. The steps to generate keys on communicating terminals are as follows:

1. Client generates the key and distributes to the gateway through a secure channel. Client and gateway generate a set of preference keys KK_i, where i=1, 2, 3….m.

   $KK_1 = h(X, X_{m\text{-}bit\text{-}shift}), KK_2 = h(X, KK_1)\ldots\ldots, KK_m = h(X, KK_{m-1})$

2. Client generates a random number s, which is later on send to gateway.

   C → G: *s*

3. Client selects two preference keys KK_{mid1}, and KK_{mid2}. KK_{mid1} is the middle key among $\{KK_1, KK_2, KK_3\ldots\ldots, KK_x\}$, where $x = s$ mod rm and rm stands for remaining number of keys in the set of KK_i. The second key KK_{mid2} is the middle key among $\{KK_1, KK_2\ldots. KK_{mid1}\}$. After this, session initialization key (SIK') is calculated.

   $SIK' = h(KK_{mid1}, KK_{mid2})$

   After generating SIK', $KK_{mid1}$and KK_{mid2} are removed from the terminals.

4. Gateway and client generate a set of session keys X_i, where i=1, 2…, n.

   $X_1 = h(SIK', X), X_2 = h(SIK', X_1)\ldots X_n = h(SIK', X_{n-1})$

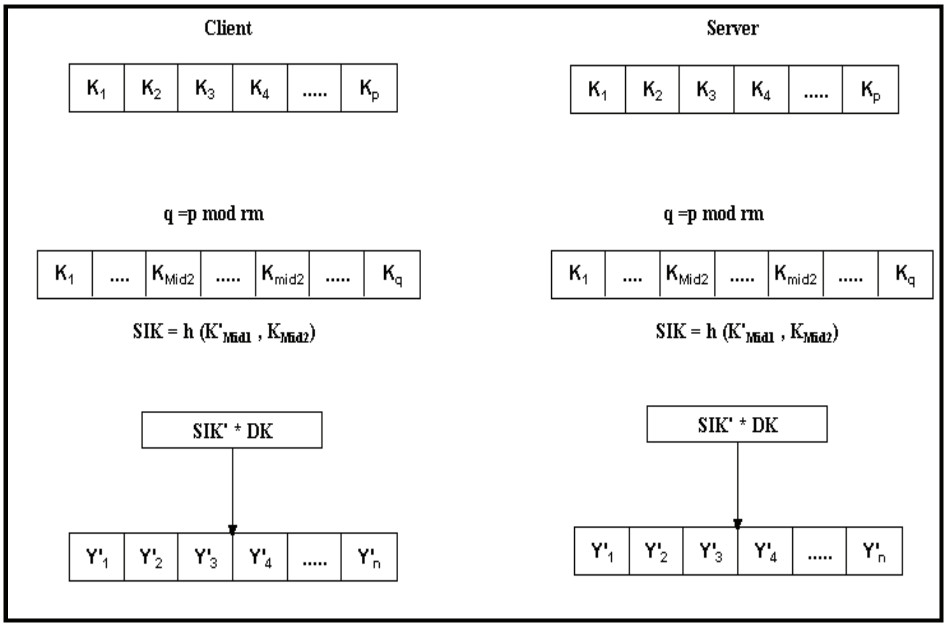
These session keys are then used by gateway and client to encrypt, decrypt and verify the messages sent amongst them.

Updating Session Keys

Figure 4 shows the procedure to update session keys. The technique proposed for updating session keys Y_i and X_i is similar, so for this section Y_i will be used for representing both. The session key update is performed as follows:

- **Step 1:** Client and the server have used Y_i up to Y_p, where $1 \le p \le n$. Then two preference keys are chosen from remaining K_i: the first key selected is K'_{Mid1}, where K'_{Mid1} is the middle key among $\{K_1\ldots K_q\}$, $q = p \bmod rm$, rm is the rm is the number of preference keys in the set of K_i. The second preference key chosen is K'_{Mid2}, where K'_{Mid2} is the middle key among $\{K_1\ldots K'_{Mid1}\}$. Then, a new session initialization key (SIK') is generated as follows:

Figure 4. Updating session keys



$SIK' = h\ (K'_{Mid1}, K'_{Mid2})$

After generating SIK', K'_{Mid1} and K'_{Mid2} are removed from both the systems.

- **Step 2:** After SIK' is generated, A new set of session key Y'_i is generated where $i = 1, 2, 3 \ldots n$, using the following method.

$Y'_1 = h\ (SIK, DK)$, $Y'_2 = h\ (SIK, Y'_1) \ldots Y'_n = h\ (SIK, Y'_n - 1)$

This results in the generation of a set of new session keys (Y'_i) without redistributing or updating the distributed key DK. DK can be used repeatedly until all the preference keys in set of K_i gets used up.

Updating Distributed Key

Updating the distributed key leads to generation a new set preference keys K'_i. After new DK' has been generated and distributed, K'_i is generated as follows: It should be noted that *CER,* which was used with DK for the first time to generate K_i, is no more stored on the client terminal. Thus a preference key K_V is selected from remaining K_i, where $v = s \bmod rm$ and s is the index of most recently used session key Y_S.

$K'_1 = h\ (DK', K_v)$, $K'_2 = h\ (DK', K'_1) \ldots\ldots$, $K_m = h\ (DK', K'_{m-1})$

After having the new set of K'_i the same procedure, as described in the third section, can be used to generate the new set of session keys.

Applying the Proposed Technique

After generating the session keys X_i and Y_i and the random numbers s and r on the client terminal, the client side protocol proceeds in the following way to connect to remote server and to get its command executed.

- **Step 1: C → G:** $ID_{C,}\ s, r, ID_G Req$
  G → C: $\{ID_G\}_{X\,i}$
- **Step 2: C → G:** $\{ID_S, CI, MAC\ [(ID_C, ID_{G,} CI), Yi]\}_{Xi,}\ MAC\ [(ID_C, ID_S), Xi_{+1}]$
- **Step 3: G → S:** $\{\{MAC\ [(IDC, ID_{G,} CI), Yi], ID_{G,} CI, r, ID_C\}\ K_S\}\ K_G^{-1}$
- **Step 4: S → G:** $\{\{\{CRes\}_{Yi}, ID_C\}\ K_G\}\ K_S^{-1}$
- **Step 5: G → C:** $\{\{CRes\}_{Yi}\}_{Xi+1}$

Command information (*CI*) and command response (*CRes*) can vary according to the server to which the client wants to connect. The technique discussed above has been implemented in a lab environment using HP iPAQ Pocket PC as client device, D-link wireless bridge as access point, and Redhat Linux machine with the gateway program and server applications accessing databases on Windows and Linux machines.

DISCUSSION

Key Management

For the remaining part of the article we will be discussing about the keys on the server side, but any discussed feature on keys will also apply on the keys with the gateway. The proposed technique requires two set of keys, K_i and Y_i, to carry out the secure communication. Both of the keys are generated on the terminals of each party but only K_i is stored on the terminals. The stored set of keys reduces the overhead of generating K_i for every new session thus enhancing the performance of client device.

Security of the Proposed Technique Against Possible Attacks

The long-time shared secret (CER) is never stored on the terminal. It is removed as soon as the first K_i is generated from it and never used again. The set of preference keys, which were used to generate *SIK*, is also removed from the terminal to eliminate any chance of regenerating a previously used *SIK*. Non-deployment of CER in the proposed technique offers flexibility in making changes to the records without giving much consideration about the network connectivity. The remainder of the section points out the other attacks and security schemes that need to be employed to keep track of the attacks and how to recover from them.

Given a situation that the session key Y_i gets compromised, even then the value of SIK still remains secure due to the fact that it is totally infeasible to perform the reverse computation on the values generated by one way hash functions. In another situation where an attacker has collected a number of session keys and tries to guess the next session key, the server can keep track of the total number of incorrectly hashed or encrypted messages and can even suspend the client's account in case the number of incorrect messages goes beyond the precise limit.

Taking into account the worst scenario, where an attacker has guessed all the right values of session keys and the server has failed to track the fake messages, the short life span of session keys can embark upon this threat in a comprehensively better way. So the compromise of session keys does not concern the participating parties in a longer run.

As session keys are a result of hash function *(SIK*DK)*, there is a possibility that the session keys can be generated if these values are compromised. If an attacker has captured all the session keys and has generated the same set of session keys, the deceit can still be traced out by the system when it finds two *MAC*s (one from a valid client and one from a fake client) hashed with same session keys. A simple solution to this situation is to update the session keys. The participating parties—who ever discovers the deceit—requests the other party to update their session keys. Thus both the parties generate new set of session keys Y_i' (as described in the third section) by generating *SIK'* which makes the compromised Y_i and *SIK* as invalid.

A successful attack on this technique means that the attacker should succeed in capturing the entire set of K_i from the device of any of the participating parties and hack the values of *r*, *p* and *DK* during distribution. Given a scenario as mentioned the deceit will still be detected by the system when it receives a *MAC* with a previously used session key.

The proposed technique inherits these security features because of the property that it never re-uses a session key and, moreover, the generation of these session keys are not based on any long term shared key. These keys are generated from randomly chosen set of preference keys K_i, which are deleted after they are used. Thus, more and more usage of the technique for conducting sessions keeps on enhancing the security of the system.

CONCLUSION

In this article we presented the wireless client model for remote access to servers by mobile clients based on a simple but secure key generation protocol. We modified and implemented the technique that was originally proposed by Kungpisdan, Le, and Srinivasan (2003) in our protocol to generate session keys for encrypting and hashing the data ought to be transmitted. Then we applied the protocol in our model to demonstrate the security provided to the wireless communications between mobile clients and servers. Our work considerably enhances the systems capability to alleviate attacks based on key compromise.

REFERENCES

Borisov, N., Goldberg, I., & Wagner, D. (2001). *Intercepting mobile communications: The insecurity of 802.11*. Paper presented at the 7th annual International Conference on Mobile Computing and Networking, Rome, Italy.

Boyd, C., & Park, D. G. (1998). Public key protocols for wireless communications. In Proceedings of the ICISC 1998 (pp. 47-57). Seoul, Korea.

Buszko, D., Lee, W. H., & Helal, A. (2001). *Decentralized ad-hoc groupware API and framework for mobile collaboration*. Paper presented at the 2001 International ACM SIG-

GROUP Conference on Supporting Group Work, Boulder, Colorado, USA.

FIPS. (1995). *Secure hash standard* (SHS). Federal Information Processing Standards Publication (FIPS) PUB 180-1.

Horn, G., & Preneel, B. (1998). Authentication and payment in future mobile systems. In *Proceedings of 5th European Symposium on Research in Computer Security* (pp. 277-293). Belgium.

IEEE. (2000). *Supplement to IEEE standard 802.11b-1999, Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Higher-speed physical layer Extension in the 2.4 GHz band.* New York. IEEE Publications.

Jing, J., Helal, A., & Elmagarmid, A. (1999). Client-server computing in mobile environments. *ACM Computing Surveys (CSUR)*, *31*(2), 117-157.

Josang, A., & Sanderud, G. (2003). *Security in mobile communications: Challenges and opportunities*. Paper presented at the Australasian Information Security Workshop Conference on ACSW Frontiers 2003, Adelaide, Australia.

Kungpisdan, S., Le, P. D., & Srinivasan, B. (2004). *A limited-use key generation scheme for Internet transactions*. In *Proceedings of the 5th International Workshop on Information Security Applications 2004 (WISA2004)*. Korea.

Kungpisdan, L., & Srinivasan, S., Srinivasan, B., & Le, P. D. (2003) *Lightweight mobile credit-card payment protocol*. In *Proceedings of the 4th International Conference on Cryptology* (pp. 295-308). Lecture Notes in Computer Science.

Pasquale, J., Hung, E., Newhouse, T., Steinberg, J., & Ramabhadran, N. (2002). *Improving wireless access to the Internet by extending the client/server model.* Paper presented at the European Wireless Conference, Florence, Italy.

Toh B. T. S., Kungpisdan, S., & Le, P. D. (2004, December 1-3) *KSL protocol: Design and implementation*. In *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems* (pp. 544-549). Singapore.

Varshney, U., & Vetter, R. (2000). Emerging mobile and wireless networks. *Communications of the ACM*, *43*(6), 73-81.

Wong, D. S., & Chan, A. H. (2001). *Efficient and mutually authentication key exchange for low power computing devices*. LNCS 2248.

Zhu, F., Wong, D. S., Chan, A. H., & Ye, R. (2002). *Password authenticated key exchange based on RSA for imbalanced wireless networks*. LNCS 2433.

KEY TERMS

Authenticated Key Exchange (AKE): AKE is block of operation in which two parties generate shared keys secretly. Many information exchanges are authenticated and secured on the basis of these shared keys.

Hash Function: Hash functions are algorithms that accept messages of any length and compute a fixed length string termed as a digital fingerprint or message digest or hash value.

Long-Term Shared Key: Long-term shared keys are keys that are shared for a longer period of time. This key forms the basis of authentication and security of information exchange between two parties.

Message Authentication Code: MAC is defined as checksum for the data transferred through unreliable media like the Internet. This check sum is also computed at the receiving ends and then the similarity of the checksum computed and received proves the authenticity of the data.

Mobile Environment: Is a generic term used to describe the environment where portable and computing devices connect wirelessly to utilize the central repository of data or applications.

Session Initialization Key (SIK): SIK can be considered as the mother of all session keys. SIK is hashed recursively with long-time shared key to generate session keys**.** One SIK expires after it has generated a predefined number of session keys.

Session Keys: Session keys are short lifespan encryption keys, which are used for encrypting one message or group of messages for a session.

Wireless Network Security

Kevin Curran
University of Ulster, Northern Ireland

Elaine Smyth
University of Ulster, Northern Ireland

INTRODUCTION

Wireless networks have a number of security issues. Signal leakage means that network communications can be picked up outside the physical boundaries of the building in which they are being operated, meaning a hacker can operate from the street outside or discretely from blocks away. In addition to signal leakage, the wired equivalent privacy protocol is inherently weak, and in addition to WEP's weaknesses, there are various other attacks that can be initiated against WLANs, all with detrimental effects. On the surface WLANs act the same as their wired counterparts, transporting data between network devices. However, there is one fundamental, and quite significant, difference: WLANs are based on radio communications technology as an alternative to structured wiring and cables. Data is transmitted between devices through the air by utilizing the radio waves. Devices that participate in a WLAN must have a network interface card (NIC) with wireless capabilities. This essentially means that the card contains a small radio device that allows it to communicate with other wireless devices within the defined range for that card, for example, the 2.4-2.4853 GHz range. For a device to participate in a wireless network, it must firstly be permitted to communicate with the devices in that network, and secondly it must be within the transmission range of the devices in that network. To communicate, radio-based devices take advantage of electromagnetic waves and their ability to be altered in such a manner that they can carry information, known as modulation (Sundaralingham, 2004). Here we discuss wireless security mechanisms.

BACKGROUND

Wired networks have always presented their own security issues, but wireless networks introduce a whole new set of rules with their own unique security vulnerabilities. Most wired security measures are just not appropriate for application within a WLAN environment; this is mostly due to the complete change in transmission medium. However, some of the security implementations developed specifically for WLANs are also not terribly strong. Indeed, this aspect could be viewed as a *work-in-progress*; new vulnerabilities are being discovered just as quickly as security measures are being released. Perhaps the issue that has received the most publicity is the major weaknesses in WEP, and more particularly the use of the RC4 algorithm and relatively short initialization vectors (IVs). WLANs suffer from all the security risks associated with their wired counterparts; however, they also introduce some unique risks of their own. The main issue with radio-based wireless networks is signal leakage. Due to the properties of radio transmissions, it is impossible to contain signals within one clearly defined area. In addition, because data is not enclosed within cable, it makes it very easy to intercept without being physically connected to the network (Hardjono & Lakshminath, 2005). This puts it outside the limits of what a user can physically control; signals can be received outside the building and even from streets away. Signal leakage may not be a huge priority when organizations are implementing their WLAN, but it can present a significant security issue, as demonstrated below. The signals that are transmitting data around an organization's office are the same signals that can also be picked up from streets away by an unknown third party. This is what makes WLANs so vulnerable. Before WLANs became common, someone wishing to gain unauthorized access to a wired network had to physically attach themselves to a cable within the building. This is why wiring closets should be kept locked and secured. Any potential hacker had to take great risks to penetrate a wired network. Today potential hackers do not have to use extreme measures, there's no need to smuggle equipment on site when it can be done from two streets away. It is not difficult for someone to obtain the necessary equipment; access can be gained in a very discrete manner from a distance.

WIRELESS SECURITY MECHANISMS

To go some way towards providing the same level of security the cable provides in wired networks, the wired equivalent protocol (WEP) was developed. WEP was designed to provide the security of a wired LAN by encryption through use of the RC4 (Rivest Code 4) algorithm. Its primary function is to safeguard against eavesdropping (sniffing), by making the data that is transmitted unreadable by a third party who does

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

not have the correct WEP key to decrypt the data. RC4 is not specific to WEP, it is a random generator, also known as a key stream generator or a stream cipher, and was developed in RSA Laboratories by Ron Rivest in 1987 (hence the name Rivest Code). It takes a relatively short input and produces a somewhat longer output, called a pseudo-random key stream. This key stream is simply added modulo two that is exclusive ORed (XOR), with the data to be transmitted, to generate what is known as ciphertext (Briere, 2005).

WEP is applied to all data above the 802.11b WLAN layers (physical and data link layers, the first two layers of the OSI reference model) to protect traffic such as transmission control protocol/Internet protocol (TCP/IP), Internet packet exchange (IPX), and hyper text transfer protocol (HTTP). It should be noted that only the frame body of data frames are encrypted, and the entire frame of other frame types are transmitted in the clear, unencrypted (Karygiannis & Owens, 2003). To add an additional integrity check, an initialization vector (IV) is used in conjunction with the secret encryption key. The IV is used to avoid encrypting multiple consecutive ciphertexts with the same key, and is usually 24 bits long. The shared key and the IV are fed into the RC4 algorithm to produce the key stream. This is XORed with the data to produce the ciphertext; the IV is then appended to the message. The IV of the incoming message is used to generate the key sequence necessary to decrypt the incoming message. The ciphertext, combined with the proper key sequence, yields the original plaintext and integrity check value (ICV) (Hardjono & Lakshminath, 2005). The decryption is verified by performing the integrity check algorithm on the recovered plaintext and comparing the output ICV to the ICV transmitted with the message. If it is in error, an indication is sent back to the sending station. The IV increases the key size, for example, a 104-bit WEP key with a 24-bit IV becomes a 128-bit RC4 key. In general, increasing the key size increases the security of a cryptographic technique. Research has shown that key sizes of greater than 80 bits make brute force[1] code breaking extremely difficult. For an 80-bit key, the number of possible keys—10^24, which puts computing power to the test; but this type of computing power is not beyond the reach of most hackers. The standard key in use today is 64 bit. However, research has shown that the WEP approach to privacy is vulnerable to certain attacks regardless of key size (Karygiannes & Owens, 2003). Although the application of WEP may stop casual sniffers, determined hackers can crack WEP keys in a busy network within a relatively short period of time.

WEP's Weaknesses

When WEP is enabled in accordance with the 802.11b standard, the network administrator must personally visit each wireless device in use and manually enter the appropriate WEP key. This may be acceptable at the installation stage of a WLAN or when a new client joins the network, but if the key becomes compromised and there is a loss of security, the key must be changed. This may not be a huge issue in a small organization with only a few users, but it can be impractical in large corporations, which typically have hundreds of users (Gavrilenko, 2004). As a consequence, potentially hundreds of users and devices could be using the same, identical key for long periods of time. All wireless network traffic from all users will be encrypted using the same key; this makes it a lot easier for someone listening to traffic to crack the key, as there are so many packets being transmitted using the same key. Unfortunately, there were no key management provisions in the original WEP protocol.

A 24-bit initialization vector WEP is also appended to the shared key. WEP uses this combined key and IV to generate the RC4 key schedule; it selects a new IV for each packet, so each packet can have a different key (Walker, 2002). Mathematically there are only 16,777,216 possible values for the IV. This may seem like a huge number, but given that it takes so many packets to transmit useful data, 16 million packets can easily go by in hours on a heavily used network. Eventually the RC4 algorithm starts using the same IVs over and over. Thus, someone passively *listening* to encrypted traffic and picking out the repeating IVs can begin to deduce what the WEP key is. Made easier by the fact that there is a static variable (the shared key), an attacker can eventually crack the WEP key (Nakhjiri, 2005). For example, a busy AP, which constantly sends 1,500 byte packets at 11Mbps, will exhaust the space of IVs after 1,500 x 8/(11 x 10^6) x 2^24 = 18,000 seconds, or 5 hours. (The amount of time may actually be smaller since many packets are less than 1,500 bytes). This allows an attacker to collect two ciphertexts that are encrypted with the same key stream. This reveals information about both messages. By XORing, two ciphertexts that use the same key stream would cause the key stream to be cancelled out and the result would be the XOR of the two plaintexts (Vines, 2002).

War-Driving

So called *war-driving* is a term used to describe a hacker who—armed with a laptop, a wireless NIC, an antenna, and sometimes a GPS device—travels, usually by car, scanning or sniffing for WLAN devices, or more specifically unprotected or *open* and easily accessed networks. The name is thought to have come from another hacking technique called war-dialing, where a hacker programs a system to call hundreds of phone numbers in search of a poorly protected computer dial-up (Nakhjiri, 2005). Due to the increased use of WLANs in recent years, it is quite possible that the number of unsecured devices has also risen in tandem, thus providing potential hackers with more choice. After all that has been written about the insecurities of WLAN, some users/organizations still insist on implementing them with their default settings and no encryption (Ulanoff, 2003). There is a plethora of hacking tools widely available to download from the Internet for any potential war-driver to use. There has been a lot of press globally, and many articles and papers written about wireless networks and their security vulnerabilities. However, despite all the literature, some enterprises still make the mistake of believing that they do not have to worry about wireless security if they are running non-critical systems with non-sensitive information across their WLANs. All information is sensitive information, and what an enterprise may class as being non-sensitive to them may be very useful to a hacker. In addition, most WLANs will connect with the wired enterprise backbone at some point, thus providing hackers with a launch pad to the entire network. The havoc an unwelcome third party could cause from here would be unlimited and very difficult to trace. Aside from the various attacks they could instigate (DoS and viruses), the loss of confidentiality, privacy, and integrity that would occur if someone where able to steal, alter, or delete information on your customer database is damaging enough. Access to sensitive information would be made relatively easy, perhaps even customers' credit card details. This could have an un-quantifiable effect on business, perhaps resulting in the loss of customers/clients and future revenue (AirDefense, 2003).

Wireless Attack Methods

A passive attack is an attack on a system that does not result in a change to the system in any way; the attack is purely to monitor or record data. Passive attacks affect confidentiality, but not necessarily authentication or integrity. *Eavesdropping* and *traffic analysis* fall under this category. When an attacker eavesdrops, he or she simply monitors transmissions for message content. It usually takes the form of someone listening into the transmissions on a LAN between stations/devices.

Eavesdropping is also known as sniffing or wireless footprinting. There are various tools available for download online which allow the monitoring of networks and their traffic; these are developed by hackers, for hackers. Netstumbler, Kismet, Airsnort, WEPCrack, and Ethereal are all well-known names in wireless hacking circles, and all are designed specifically for use on wireless networks, with the exception of Ethereal, which is a packet analyzer and can also be used on a wired LAN. NetStumbler and Kismet can be used purely for passive eavesdropping; they have no additional active functions, except perhaps their ability to work in conjunction with global positioning systems (GPSs) to map the exact locations of identified wireless LANs. NetStumbler is a Windows-based sniffer, where Kismet is primarily a Linux-based tool. NetStumbler uses an 802.11 Probe Request sent to the broadcast destination address, which causes all APs in the area to issue an 802.11 Probe Response containing network configuration information, such as their SSID, WEP status, the MAC address of the device, name (if applicable), the channel the device is transmitting on, the vendor and the type, either peer or AP, along with a few other pieces of information. Using the network information and GPS data collected, it is then possible to create maps with tools such as StumbVerter and MS Mappoint.

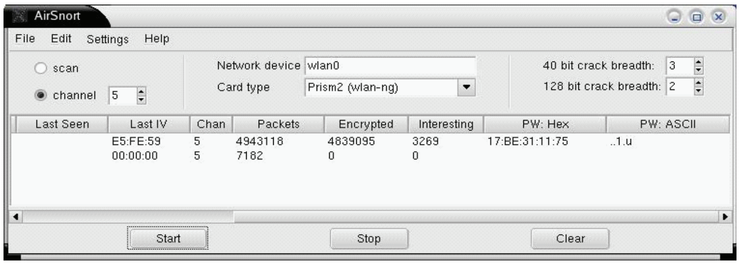
Kismet, although not as graphical or user friendly as NetStumbler, is similar to its Windows counterpart, but it provides superior functionality. While scanning for APs, packets can also be logged for later analysis. Logging features allow for captured packets to be stored in separate categories, depending upon the type of traffic captured. Kismet can even store encrypted packets that use weak keys separately to run them through a WEP key cracker after capture, such as Airsnort or WEPCrack (Sundaralingham, 2005). Wireless network GPS information can be uploaded to a site called Wigle (http://www.wigle.net). Therefore, if Wigle data exists for a particular area, there is no need to drive around that area probing for wireless devices; this information can be obtained in advance from the Wigle Web site. All that remains is to drive to a location where known networks exist to observe traffic. Wigle currently has a few hundred thousand networks on its database.

Traffic analysis gains intelligence in a more subtle way by monitoring transmissions for patterns of communication. A considerable amount of information is contained in the flow of messages between communicating parties. Airopeek NX, a commercial 802.11 monitoring and analysis tool for Windows, analyzes transmissions and provides a useful node view, which groups detected stations and devices by their MAC address and will also show IP addresses and protocols observed for each. The Peer Map view, within Airopeek NX, presents a matrix of all hosts discovered on the network by their connections to each other. This can make it very easy to visualize AP and client relationships, which could be useful to hackers in deciding where to try and gain access or target for an attack (McClure, Scambray, & Jurtz, 2003). Some attacks may begin as passive, but then crossover to active as they progress. For example, tools such as Airsnort or WEPCrack may passively monitor transmissions, but their intent is to crack the WEP key used to encrypt data being transmitted. Figure 1 shows a screen shot of Airsnort showing the number of packets gathered and the WEP key—17BE311175.

Ultimately the reasons for wanting to crack the key are so that an unauthorized individual can access a protected network and then launch an active attack of some form or another. These types of attacks are classed as passive decryption attacks.

An active attack, also referred to as a malicious attack, occurs when an unauthorized third party gains access to a network and proceeds to perform denial of service (DoS) attack, to disrupt the proper operation of a network, to intercept network traffic and either modify or delete it, or

Figure 1. Screen shot from Airsnort showing 64-bit key crack



○ scan

● channel 5

Network device wlan0

Card type Prism2 (wlan-ng)

40 bit crack breadth: 3

128 bit crack breadth: 2

Last Seen	Last IV	Chan	Packets	Encrypted	Interesting	PW: Hex	PW: ASCII
	E5:FE:59	5	4943118	4839095	3269	17:BE:31:11:75	..1.u
	00:00:00	5	7182	0	0		

inject extra traffic onto the network. There are many active attacks that can be launched against wireless networks; the following few paragraphs outline almost all of these attacks, how they work, and what effect they have (Karygiannis & Owens, 2003). DoS attacks are easily the most prevalent type of attack against 802.11 networks and can be waged against a single client or an entire WLAN. In this type of attack, the hacker usually does not steal information; he or she simply prevents users from accessing network services, or causes services to be interrupted or delayed. Consequences can range from a measurable reduction in performance to the complete failure of the system. Some common DoS attacks are outlined below.

A man-in-the-middle attack is carried out by inserting a malicious station between the victim station and the AP, thus the attacker becomes the man in the middle; the station is tricked into believing that the attacker is the AP, and the AP into believing that the attacker is the legitimate station. To begin the attack, the perpetrator passively monitors the frames sent back and forth between the station and the AP during the initial association process with an 802.11 analyzer. As a result, information is obtained about both the station and the AP, such as the MAC and IP address of both devices, association ID for the station, and SSID of the network. With this information a rogue station/AP can be set up between the two unsuspecting devices. Because the original 802.11 does not provide mutual authentication, a station will happily re-associate with the rogue AP. The rogue AP will then capture traffic from unsuspecting users; this of course can expose information such as user names and passwords (Gavrilenko, 2004).

An Association flood is a resource starvation attack. When a station associates with an AP, the AP issues an associate identification (AID) number to the station in the range of 1-2007. This value is used for communicating power management information to a station that has been in a power-save state. This attack works by sending multiple authentication and association requests to the AP, each with a unique source MAC address. The AP is unable to differentiate the authentication requests generated by an attacker and those created by legitimate clients, so it is forced to process each request. Eventually, the AP will run out of AIDs to allocate and will be forced to de-associate stations to reuse previously allocated AIDs. In practice, many APs will restart after a few minutes of authentication flooding, however this attack is effective in bringing down entire networks or network segments; if repeatedly carried out, it can cause a noticeable decrease in network up time (Wright, 2003). The final issue is a threat posed by the simple network management protocol (SNMP).

Attacks that Alter Transmissions

The following attacks describe how it is possible for an attacker to modify messages in transit, without detection. Message modification attacks are made relatively trivial if no message encryption exists; however, even if it does, the hacker can still get around it by first cracking the encryption and then carrying out the attack.

- **Injecting Traffic:** If an attacker knows the exact plaintext for one encrypted message, he or she can then use this knowledge to construct more correctly encrypted packets. This procedure involves constructing a new message, calculating the CRC-32 checksum, and performing bit-flips[2] on the original encrypted message to change the plaintext to the new message. This packet can now be sent to the AP, and it will be accepted as a valid packet. Because RC4 encrypts data a byte at a time, an attacker can modify one byte of ciphertext and the recipient would not know the data has been changed. RC4 does not detect errors (Borisov, Goldberg, & Wagner, 2003).
- **IP Redirection:** By intercepting and modifying the IP address of the destination in a packet, an attacker can effectively re-route messages. This attack can be used where an AP acts as an IP router with Internet connectivity, which is fairly common. The idea is to take an encrypted packet that has been transmitted and modify it so it has a new destination address—one the attacker controls. The AP will then decrypt the packet and send it off to its new destination, where the attacker

can read the packet, now in the clear (Borisov et al., 2003).

- **SNMP Attack:** The final issue is a threat posed by the simple network management protocol (SNMP). Some APs can be managed via wireless link, usually with a proprietary application, replying on SNMP. Executing these operations can represent a frightening vulnerability for the whole LAN; because eavesdroppers can decipher the password to access read/write mode on the AP using a packet analyzer, this means that they share the same administration privileges with the WLAN administrator and can manage the WLAN in a malicious manner (Me, 2003). The sheer number of attacks, and their effects, would seem to put WLANs at a severe disadvantage over their wired counterparts. However, there are just as many, if not more, security measures that users can utilize to counteract most of the above attacks. Layering one security measure on top of another, to strengthen the overall system to deter any potential attackers or make their task more difficult, is not impossible. However, not all organizations, or indeed individuals, take the time to implement any form of security, or they implement it weakly. The next article discusses, firstly, what route the primary research will take, and secondly, the findings resulting from the primary research.

FUTURE TRENDS

IEEE specifies basically two categories of WLAN standards—those that specify the fundamental protocols for the complete wireless system and those that address specific weaknesses or provide additional functionality (3COM, 2006). Here we mention just three of the latter standards which may have a significant influence in the coming days.

802.11i is a major extension because it was intended to improve WLAN security on 802.11a and 802.11b networks, which was in tatters. It adds two main blocks of improvements: improved security for data in transit, and better control of who can use a network. It covers key management and distribution, encryption, and authentication, the three main components of security (Briere, 2005). The 802.11i specification can be viewed as consisting of three main sections, organized into two layers. On the lower level are improved encryption algorithms in the form of the temporal key integrity protocol (TKIP) and the counter mode with cipher block chaining-message authentication code, CBC-MAC protocol (CCMP). Both of these provide enhanced data integrity over WEP, with TKIP being targeted at legacy equipment and CCMP being targeted at future WLAN equipment (Nakhjiri, 2005).

The goal of the 802.11k standard is to make measurements from layers one and two of the OSI protocol stack—physical and data link layers—available to the upper layers. It is expected that the upper layers will then be able to make decisions about the radio environment. It is called radio resource management. One feature is better traffic distribution. Normally a wireless device will connect to whatever AP gives it the strongest signal. However, this can lead to an overload on some APs and under-load on others, resulting in an overall lowered service level. The 802.11k standard will allow network management software to detect this situation and redirect some of the users to under-utilized APs. 802.11n is a high-performance standard that would boost both 802.11b and 802.11a 11Mbps and 54Mbps, respectively. Proposals say it could go to 108Mbps or beyond, to as much as 320Mpbs. This standard is not expected to be complete until early 2007.

Community 802.11b networks will continue to grow as people realize they can share their high-speed, high-cost Internet connections, turning them into high-speed, low- or no-cost connections for a larger group of people (Imai, 2005).

CONCLUSION

Wireless networks have a number of security issues. Signal leakage means that network communications can be picked up outside the physical boundaries of the building in which they are being operated, meaning a hacker can operate from the street outside or discretely from blocks away. In addition to signal leakage, wireless networks have various other weaknesses. WEP, the protocol used within WLANs to provide the equivalent security of wired networks, is inherently weak. The use of the RC4 algorithm and weak IVs makes WEP a vulnerable security measure. In addition to WEP's weaknesses, there are various other attacks that can be initiated against WLANs, all with detrimental effects.

REFERENCES

AirDefense. (2003) *Wireless LAN security: What hackers know that you don't.* Retrieved from http://ssl.salesforce.com/servlet.Email/AttachmentDownload?q=00m0000000003Pr00D00000000hiyd00500000005k8d5

Borisov, N. Goldberg, I., & Wagner, D. (2003) *Security of the WEP algorithm.* Retrieved from http://www.isaac.cs.berkeley.edu/isaac/wep-faq.html

Briere, D. (2005, October) *Wireless network hacks and mods for dummies (for dummies S.).* New York: Hungry Minds.

Gavrilenko, K. (2004, June) *WI-FOO: The secrets of wireless hacking.* Boston: Addison-Wesley.

Hardjono, T., & Lakshminath R. D. (2005, July). *Security in wireless LANs and MANs.* Norwood, MA: Artech House.

Harte, L., Kellog, S., Dreher, R., & Schaffinit, T. (2000) *The comprehensive guide to wireless technologies: Cellular, PCS, paging, SMR and satellite.* NC: APDG.

Imai, H. (Ed.). (2005, December). *Wireless communications security.* Norwood, MA: Artech House.

Karygiannis, T., & Owens, L. (2003). *National Institute of Standards and Technology, special publication 800-48, draft.* Retrieved from http://csrc.nist.gov/publications/drafts/draft-sp800-48.pdf

McClure, S., Scambray, J., & Jurtz, G. (2003). *Hacking exposed: Network security secrets and solutions* (4th ed.). New York: Osbourne McGraw-Hill.

Me, G. (2003). *A threat posed by SNMP use over WLAN.* Retrieved from http://www.wi-fitechnology.com/Wi-Fi_Reports_and_Papers/SNMP_use_over_WLAN.html

Nakhjiri, M. (2005, September). *AAA and network security for mobile access: Radius, diameter, EAP, PKI and IP mobility*. New York: John Wiley & Sons.

Sundaralingham, S. (2004, November). *Cisco wireless LAN security.* San Francisco: Cisco Press.

3COM. (2006). Retrieved from http://www.3com.com/whitepapers.html

Ulanoff, L. (2003). Get free Wi-Fi, while its hot. *PC Magazine,* (July).

Vines, R. D. (2002). *Wireless security essentials, defending mobile systems from data piracy.* London: John Wiley & Sons.

Walker, J. (2002) *Unsafe at any key size; an analysis of the WEP encapsulation.* Retrieved from http://www.dis.org/wl/pdf/unsafe.pdf

Wright, J. (2003). *Detecting wireless LAN MAC address spoofing.* Retrieved from http://home.jwu.edu/jwright/papers/wlan-mac-spoof.pdf

KEY TERMS

Denial of Service: An incident in which a user or organization is deprived of the services of a resource they would normally expect to have. Typically, the loss of service is the inability of a particular network service to be available or the temporary loss of all network connectivity and services.

Direct Sequence Spread Spectrum (DSSS): Combines a data signal with a higher data rate bit sequence, referred to as a chipping code. The data is exclusive ORed (XOR) with a PRS that results in a higher bit rate, This increases the signal's resistance to interference.

Frequency Hopping Spread Spectrum (FHSS): Here the signal hops from frequency to frequency over a wide band of frequencies. The transmitter and receiver change the frequency they operate on in accordance with a pseudo-random sequence (PRS) of numbers. To properly communicate, both devices must be set to the same hopping code.

IEEE 802.11 Standard: IEEE has developed several specifications for WLAN technology, the names of which resemble the alphabet. There are basically two categories of standards: those that specify the fundamental protocols for the complete wireless system—these are called 802.11a, 802.11b, and 802.11g; and those that address specific weaknesses or provide additional functionality—these are 802.11d, e, f, h, I, j, k, m, and n.

Initialization Vector (IV): Used in conjunction with the secret encryption key in WEP (see *Wired Equivalent Privacy*). The IV is used to avoid encrypting multiple consecutive ciphertexts with the same key, and is usually 24 bits long.

War Driving: A term used to describe a hacker who—armed with a laptop, a wireless NIC, an antenna, and sometimes a GPS device—travels, usually by car, scanning or sniffing for WLAN devices, or more specifically unprotected or open and easily accessed networks.

Wired Equivalent Privacy (WEP): Designed to provide the security of a wired LAN by encryption through use of the RC4 (Rivest Code 4) algorithm. Its primary function is to safeguard against eavesdropping (sniffing), by making the data that is transmitted unreadable by a third party who does not have the correct WEP key to decrypt the data.

ENDNOTES

[1] A method that relies on sheer computing power to try all possibilities until the solution to a problem is found; usually refers to cracking passwords by trying every possible combination of a particular key space.

[2] Bit-flipping—changing one or more bits within a message. For example, change a 0 to a 1, or vice versa.

Wireless Security

Meletis Belsis
Telecron, Greece

Alkis Simitsis
National Technical University of Athens, Greece

Stefanos Gritzalis
University of the Aegean, Greece

INTRODUCTION

The fast growth of wireless technology has exponentially increased the abilities and possibilities of computing equipment. Corporate users can now move around enterprise buildings with their laptops, PDAs, and WiFi; enable VoIP handsets; and retain communications with their offices. Business users can work from almost anywhere by attaching their laptops to WiFi hotspots and connecting to their corporate network. However, not many enterprises know and understand the potential security vulnerabilities that are introduced by the use of WiFi technologies. Wireless technologies are insecure by their nature. Anyone with the appropriate hardware can steal information transmitted using the airwaves. This article discusses the security vulnerabilities that are inherited in wireless networks. Also, it provides a description of the current security trends and protocols used to secure such WiFi networks along with the problems from their application.

BACKGROUND

Currently, several enterprises consider information security as a monolithic architecture, in which simply they install a firewall or an intrusion detection system. Unfortunately security is not a single device or software:

> *In the real world, security involves processes. It involves preventive technologies, but also detection and reaction processes, and an entire forensics system to hunt down and prosecute the guilty. Security is not a product; it itself is a process.* (Schneier, 2000)

The above definition represents the fact that total protection of corporate networks goes beyond a firewall engine. Each appliance that is added and/or changed into a system should incorporate the re-designing of a system's overall security policy and infrastructure. The same principle exists when incorporating wireless devices to extend the overall enterprise architecture. Deploying a wireless network has as a consequence the change of the security risks and needs of the entire network infrastructure. Nowadays, the techniques that are used for the realization of attacks in wireless connected networks resemble those used to target common LANs. In the next paragraphs, we present the major categories of attacks, including techniques that have been successfully used for attacking corporate wireless networks.

Denial of Service. In their simplest form, an adversary can continuously transmit *association request* packets. Such action could render an access point unavailable to authorized users. Adversaries can use a powerful RF transceiver to transmit amplified signals in all frequency band frequencies (channels), creating an interjection that prevents the communication of terminals with the corporate access points (RF Jamming). Such an attack could be easily deployed from the outside premises of an enterprise (e.g., parking). An example appliance that can be used for the concretization of this attack is the Power Signal Generator (PSG -1) by the YDI.

Man-in-the-Middle Attacks. Combining an RF Jamming attack with the use of a portable computer and necessary software, an attacker can easily steal or alter corporate information (Akin, 2003). The adversary will use a denial-of-service attack to force authorized terminals connected to a corporate access point to identify and roam to an access point with better signal than the one already connected to. Using this predetermined behavior the attacker can masquerade his/her laptop as an access point and force all wireless clients to connect to it. By using this technique an adversary can intercept all wireless communications links and read or alter information on them.

Fresnel Zone Sniffing. Stealing information from point-to-point wireless links is difficult. The attacker needs to calculate the link path and identify ways to attach its laptop to the link's Fresnel Zone.

Rogue Wireless Gateways. A rogue wireless gateway is a security vulnerability that is detected in many of today's

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

enterprise networks. A rogue wireless gateway is an unauthorized access point that is installed on an enterprise network. Such access points are usually installed by corporate users, to assist them in the everyday work (i.e., transfer files/e-mails from a desktop to a laptop computer). Unfortunately enterprise users do not know and understand the security implications of installing a wireless device on a system. Leaving such devices connected to the corporate network provides an opportunity to adversaries to connect and steal corporate information.

Ad Hoc Networks. The 802.11 protocol specification allows wireless terminals to interconnect without the use of an access point. This mode of operation is called ad hoc. Unfortunately many of today's corporate users enable the ad hoc facility on their laptops and PDAs either accidentally or deliberately in order to exchange files with other users. Enabling the ad hoc mode without deploying the necessary security procedures (i.e., encryption and authentication) could seriously damage corporate security. Adversaries can search for such unprotected ad hoc networks and connect to those. From there adversaries can either read the locally stored corporate information, or if the user's device is connected to the corporate networks (i.e., LAN, dialup, and VPN), access the corporate resources (Papadimitratos & Haas, 2002).

The previous example attacks emphasize the need for security that results from the use of wireless technology. The problem of security becomes more apparent when the technology of wireless networking is applied in government-owned systems. The need for security in those systems is extensive due to the legislation on personal data protection and the human lives factors involved.

MAIN THRUST OF THE ARTICLE

In the last few years, the computing and telecommunications community has realized the necessity of deploying security controls on wireless networks. Unfortunately most of today's wireless security controls have been proven unsafe or managerial infeasible to maintain. The next few paragraphs describe the most common security protocols and techniques, as well as their vulnerabilities.

Discovering Wireless Networks

Many enterprises support their notion of using insecure WiFi networks based on the idea that their *small wireless networks* are hidden from hackers and adversaries. This notion is called Security through Obscurity, and is something that the IT security community analyzed and abolished long before the appearance of wireless networks.

Modern hackers have invented a number of new techniques, collectively known as *War Driving* or *War Chalking,* which aim at discovering unprotected wireless networks. An adversary uses a laptop computer, along with appropriate discovery software (i.e., NetStumbler) and a GPS received, to pinpoint the exact location of access points on a map. Today such maps are distributed among the War Driving community. It is not unusual for enterprises to discover their company access points on maps found on War Driving Web sites (see Figure 1).

Many enterprise administrators try to hide their wireless networks by activating the *close system* option found on access point hardware equipment. This option prohibits the access point from transmitting the network's beacon information that incorporates the network's service set identifier (SSID). Unfortunately the SSID is incorporated into almost all network management frames. Software packages like

Figure 1. A War Driving result in Los Angeles

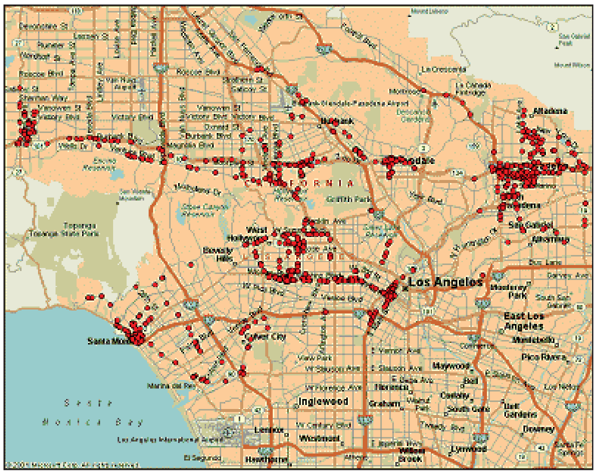
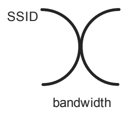
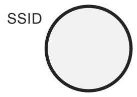
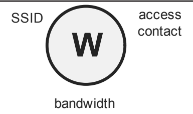


Table 1. War Chalking symbols

node	symbol
open node	
closed node	
WEP node	

NetStumbler will force the access points in transmitting the SSID by issuing such management frames (i.e., *Reassociation Request*).

The techniques of War Driving and War Chalking has been used today to an extended degree, and adversaries have developed their own marking symbols (see Table 1) in order to denote the buildings where wireless networks are discovered. Writing these symbols in various buildings of the city, adversaries mark their potential targets.

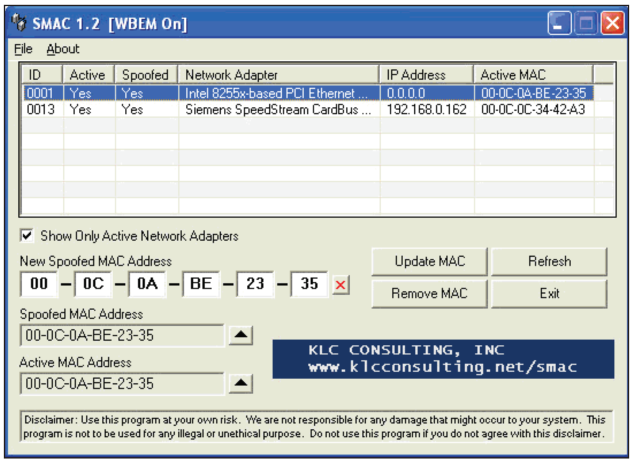
MAC Access Control Lists

To enhance security, many corporations develop media access control (MAC) *control lists* declaring the MAC addresses of wireless terminals that are authorized to access the wired segment of a corporate network. Unfortunately the deployment of MAC access control lists increases management time and difficulty without offering real protection from experienced hackers. Having discovered a wireless network, an adversary can eavesdrop on the network and detect authorized MAC addresses that connect to an access point. Having a list of such authorized MAC addresses, the adversary can use MAC spoofing attacks and masquerade his laptop as an authorized client (e.g., using the SMAC software, a snapshot of which is depicted in Figure 2).

Wired Equivalent Privacy (WEP)

The first security protocol developed for wireless networks is *wired equivalent privacy* (WEP). WEP uses RC4 PRNG algorithm (LAN MAN, 1999) for the coding of information. The WEP key, with a 24-bit *initializing vector* (IV), is used for the encryption/decryption of wireless data. The protocol works with keys of 64 or with 128 bit (the actual key lengths are 40 and 104 bit, but are concatenated with the IV during the encryption phase). In a WEP environment the encryption keys are installed by the administrator of the system in each terminal and access point, and thus the management of the network becomes more complicated.

Figure 2. SMAC software screenshot



ID	Active	Spoofed	Network Adapter	IP Address	Active MAC
0001	Yes	Yes	Intel 8255x-based PCI Ethernet ...	0.0.0.0	00-0C-0A-BE-23-35
0013	Yes	Yes	Siemens SpeedStream CardBus ...	192.168.0.162	00-0C-0C-34-42-A3

☑ Show Only Active Network Adapters

New Spoofed MAC Address: 00 – 0C – 0A – BE – 23 – 35

Spoofed MAC Address: 00-0C-0A-BE-23-35

Active MAC Address: 00-0C-0A-BE-23-35

Disclaimer: Use this program at your own risk. We are not responsible for any damage that might occur to your system. This program is not to be used for any illegal or unethical purpose. Do not use this program if you do not agree with this disclaimer.

The WEP does not offer user authentication; therefore, discovering the WEP key allows access to a corporate network (Borisov, Goldberg, & Wagner, 2001). The two authentication models provided by WEP are open system and the shared-key authentication (Lambrinoudakis & Gritzalis, 2005). The open system model uses the MAC access control lists discussed in the previous paragraphs. In the shared key authentication, WEP uses the encryption key to implement a Challenge-Response authentication scheme.

At the same time WEP uses a 32-bit *cycle redundancy check* algorithm as integrity check value (ICV) in order to ensure the integrity of data. Currently, the CRC algorithm has already been broken by researchers from the University of Berkley (Tyrrell, 2003).

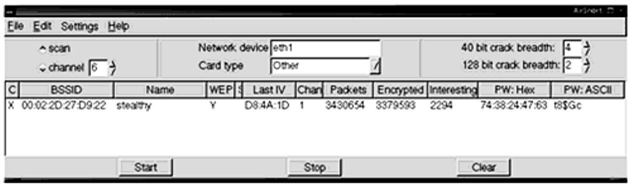
The key recovery process in a system that uses WEP can be actually realized in a few hours. This is due to a vulnerability found in the way WEP uses the RC4 algorithm. The weakness of WEP is based on the fact that the IV is only 24 bit, and thus, in a busy network, the same IV key is used to encrypt different network packets. Having eavesdropped two or more packets encrypted with the same IV, an adversary can apply cryptanalysis techniques and recover the WEP key. Today, a number of freeware software packages that can perform a successful WEP attack are available on the Internet. Examples of such software artifacts include WEPCrack and AIRSnort (see Figure 3).

Due to the fact that WEP encryption keys are static, the time between discovering a compromised key and updating the whole wireless network infrastructure with a new key is extended. This leaves even more time for adversaries to access and copy confidential corporate information.

WiFi-Protected Access (WPA)

Understanding the problems of WEP, the international community has moved forward in developing a more secure protocol, namely 802.11i (Edney & William, 2003). Due to the delay in the development of the final 802.11i standard, the international community released a pre-802.11i security

Figure 3. AirSnort software screenshot



○ scan
● channel 6

Network device: eth1

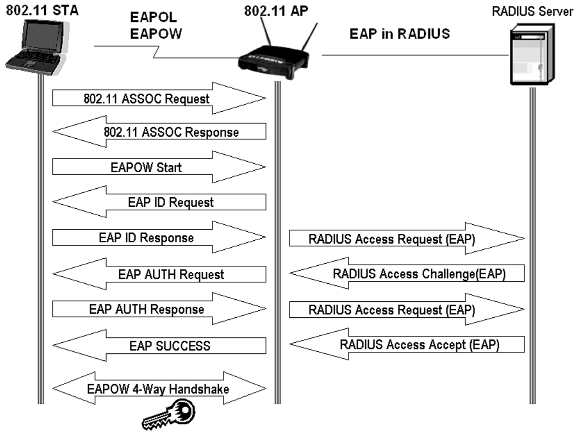
Card type: Other

40 bit crack breadth: 4

128 bit crack breadth: 2

C	BSSID	Name	WEP	S	Last IV	Chan	Packets	Encrypted	Interesting	PW: Hex	PW: ASCII
X	00:02:2D:27:D9:22	stealthy	Y		D8:4A:1D	1	3430654	3379593	2294	74:38:24:47:63	t8$Gc

Figure 4. 802.1x EAP authentication (EAP Authentication, 2005)



protocol under the name *WiFi-protected access* (Edney & William, 2003).

The WPA uses algorithm RC 4 (Fluhrer, Mantin, & Shamir, 2001) for the encryption of air data incorporating the *temporal key integrity protocol* (TKIP), in order to use dynamic encryption keys. In order to avoid the security vulnerabilities of CRC-32, WPA utilizes a novel integrity protection algorithm, the Michael Message Integrity Check (MIC) (Cam-Winget, Housley, Wagner, & Walker, 2003), which uses a 64-bit key and partitions data into 32-bit blocks.

TKIP uses an IV of 48 bit, offering better security than the 24-bit IV used by WEP. It combines a 128-bit temporary key, which is preinstalled in all wireless terminals with the MAC address of each terminal, and the 48-bit IV in order to create a new encryption key for each terminal. The protocol changes the encryption key every 10,000 packets that are transmitted.

Moreover, WPA employs the 802.1x protocol (port-based access control) to deliver authenticated connections. This protocol allows the usage of a number of authentication methods to be used such as passwords and digital certificates.

The user or terminal authentication process is performed by the *extensible authentication protocol* (EAP), which is usually associated with a radius server in order to securely authenticate users or devices on a network. Figure 4 displays an example EAP authentication process.

Currently, there exist several EAP implementations:

- **EAP-MD 5** (Funk, 2003): This was the first protocol to use user authentication based on the 802.1x scheme. It provides only one-way authentication, ensuring the authenticity of users but not the servers. The protocol is based on the algorithm MD5. However, research has already proven that this protocol is subject to dictionary and man-in-the-middle attacks (Asokan, Niemi, & Nyberg, 2002).
- **CISCO-LEAP:** The lightweight EAP (LEAP), created by CISCO, offers bidirectional authentication. The bidirectional authentication makes the protocol immune to man-in-the-middle attacks, but its challenge handshake authentication protocol (MSCHAP version 2) is subject to dictionary attacks. Currently, there exist several tools on the Internet, like the *asleep,* that can perform successful attacks on LEAP. CISCO tries to tackle this disadvantage, and at this time, the company is developing a new protocol called EAP-FAST.
- **EAP-FAST** (Ghosh & Gupta, 2005): EAP-FAST, developed and marketed by CISCO, is thought to be as secure as EAP-PEAP and as easy to deploy as EAP-LEAP. The protocol operates similarly to EAP-PEAP. It uses two distinct phases. In phase 1 a secure tunnel is established using a Protected Access Credential (PAC) shared key. PAC is used in order to avoid deploying digital certificates. After the establishment of the secure tunnel, authentication is performed on phase 2 using the MSCHAP v2 protocol. The PAC secret can either be manually shared to all nodes or automated through an optional Diffie-Hellman process. Unfortunately, using the manual shared key distribution process will make the management of the network extremely difficult. On the other hand the anonymous Diffie-Hellman process can make the protocol suspect to man-in-the-middle attacks. Along with this, during the anonymous Diffie-Hellman, the protocol transmits the user name in cleartext (unencrypted), and thus possession of a user name could further lead an attacker to perform social engineering attacks. It is going to be a while before the protocol is thoroughly tested and used by the international community (Lambrinoudakis & Gritzalis, 2005).
- **EAP-TLS** (Aboba & Simon, 1999): The EAP-transport layer security (EAP-TLS), developed by Microsoft Corporation, uses the Transport Layer Security (TLS) protocol with digital certificates for both clients and servers in order to provide bidirectional authentication. The protocol transmits the user name in cleartext. A possible information leakage in this form could provide the basis for further attacks (i.e., social engineering). Along with this, the use of both client and server certificates makes the management of this protocol a hassle for large corporate networks.
- **EAP-TTLS** (Funk & Blake-Wilson, 2003): The EAP-tunneled TLS (EAP-TTLS) protocol, created by the companies *Funk* and Certicom, is based on the idea of EAP-TLS, but in order to minimize the management process, it uses their digital certificates only for the servers and not for the clients. Clients authenticate servers by using digital certificates; thus, the proto-

col builds an *encrypted tunnel*. The encrypted tunnel provides a secure medium on which clients can be authenticated using a challenge response mechanism. Although, currently, there are no known attacks, the protocol is suspected to be vulnerable to man-in-the-middle attacks (Asokan et al., 2002).

- **EAP-PEAP** (Palekar et al., 2003): The protected EAP (PEAP) protocol is the result of a common effort from different IT companies. PEAP uses digital certificates for servers. Also, clients authenticate servers. After a successful server authentication, the protocol creates an encrypted tunnel between the client and the server. Inside this secure tunnel, the system can use any of the previously described EAP authentication methods in order to enable client authentication. The chosen combination today is to use the EAP-TLS inside the encrypted tunnel in order to provide client authentication (EAP-PEAP/EAP-TLS). Similar to the TTLS protocol, no known attack exists today, but PEAP is suspected to be vulnerable to man-in-the-middle attacks.

802.11I

Having discovered the vulnerabilities in WEP, IEEE started producing the specification of a new protocol, IEEE 802.11i. It follows principles similar to WPA, and uses 802.1x and EAP protocols for authentication and key management. 802.11i uses the *counter-mode/CBC-MAC protocol* (CCMP) protocol with the *advance encryption standard* (AES) (NIST, 2001) algorithm to provide data encryption and integrity protection.

In addition, 802.11i provides the robust security network (RSN) feature. RSN allows the two ends of a communication link to negotiate the encryption algorithms and protocols to be used. This facility enables updating a wireless network with new algorithms and protocols in order to protect it from future vulnerabilities.

Still, the 802.11i protocol requires special encryption hardware to run the AES algorithm; due to this fact, additional time is needed for the vendors to change their existing hardware to support the 802.11i protocol. To enable the migration of WEP and WPA systems to 802.11i, the WiFi Alliance has proposed a new security protocol—the WPA2. The new protocol incorporates all 802.11i functionality, but also enables the use of the TKIP protocol to support devices that do not have the necessary hardware to run the AES algorithm.

VPNs

To provide a solution to the problem of security, many companies are extending/developing *virtual private networks* (VPNs) (Karygiannis & Owens, 2002). Maintaining a VPN requires the engagement of specialized personnel or the training of existing personnel; in both cases, the costs associated with deploying a wireless infrastructure are highly increased. Along with the cost associated with the deployment of a VPN, VPNs incorporate a number of operational problems on a system.

In networks where the users roam contentiously, a Layer-3 VPN solution will disrupt a user's connection and may even force the user to re-authenticate. Along with this, applications that run on client terminals and access data stored on the corporate servers may be seriously disrupted from a Layer-3 disconnection. Such disconnections can seriously damage the integrity and availability of corporate information.

CONCLUSION

In this article, we have discussed the critical issue of wireless security. We have presented the security vulnerabilities that are frequently inherited in wireless networks. Also, we have described the most common security protocols and techniques used. Moreover, we have provided a description of the current security trends and protocols used to secure such WiFi networks, along with the problems from their application.

REFERENCES

Aboba, B., & Simon, D. (1999). *PPP EAP TLS authentication protocol.* IETF RFC 2716.

Akin, D. (2003). *Certified Wireless Security Professional (CWSP) official study guide.* New York: McGraw-Hill.

Asokan, N., Niemi, V., & Nyberg, K. (2002). *Man-in-the-middle in tunneled authentication protocols.* Cryptology ePrint Archive, Report 2002/163.

Borisov, N., Goldberg, I., & Wagner, D. (2001). *Intercepting mobile communications: The insecurity of 802.11.* Retrieved December 16, 2005, from http://www.isaac.cs.berkeley.edu/isaac/mobicom.pdf

Cam-Winget, N., Housley, H., Wagner, D., & Walker, J. (2003). Security flaws in 802.11 data link protocols. *Communications of the ACM, 46*(5).

EAP Authentication. (2005). Retrieved December 13, 2005, from http://www.wi-fiplanet.com

Edney, J., & William, A. (2003). *Real 802.11 security: WiFi protected access and 802.11i.* Boston: Addison-Wesley.

Fluhrer, S., Mantin, I., & Shamir, A. (2001). Weaknesses in the key scheduling algorithm of RC4. *Proceedings of the 8th Annual Workshop on Selected Areas in Cryptography.* Berlin: Springer-Verlag (LNCS 2259).

Funk, P. (2003). *The EAP MD5-Tunneled authentication protocol (EAP-MD5-Tunneled).* IETF Internet Draft.

Funk, P., & Blake-Wilson, S. (2003). *EAP Tunneled TLS authentication protocol (EAP-TTLS).* IETF Internet Draft.

Ghosh, D., & Gupta, A. (2005). *Analysis of EAP-FAST wireless security protocol.* Retrieved December 15, 2005, from http://wwwcsif.cs.ucdavis.edu/~guptaa/finalreport.pdf

Karygiannis, T., & Owens, L. (2002). *Wireless network security.* NIST Special Publication 800-48.

Lambrinoudakis, C., & Gritzalis, S. (2005). *Security in IEEE 802.11 WLANS.* CRC Press.

LAN MAN. (1999). *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. IEEE Standard 802.11, 1999 edition.* Standards Committee of the IEEE Computer Society.

NIST. (2001). *Announcing the Advance Encryption Standard (AES).* Federal Information Processing Standards Publication 197.

Palekar, A., Simon, D., Zorn, G., Salowey, J., Zhou, H., & Josefsson, S. (2003). *Protected EAP Protocol (PEAP) version 2.* IETF Internet Draft.

Papadimitratos, P., & Haas, Z.J. (2002). Secure routing for mobile ad hoc networks. Working session on security in wireless ad hoc networks, EPFL. *Mobile Computing and Communications Review, 6*(4).

Schneier, B. (2000). *Secret and lies* (1st ed.). New York: John Wiley & Sons.

Tyrrell, K. (2003). *An overview of wireless security issues.* SANS Information Security Reading Room, SANS Institute.

KEY TERMS

Encrypted Tunnel: An encrypted logical (virtual) connection between two ends. Data traveling inside the tunnel are encrypted with an agreed encryption algorithm.

Fresnel Zone: The area around the visual line of sight of a wireless link on which the RF waves are spread. This area must be clear from obstacles, otherwise the RF signal is weakened.

Man-in-the-Middle Attack: An attack where the adversary succeeds in locating himself in an established connection between two or more authorized nodes. Data traveling between the nodes are always passing from the adversary.

Reassociation Request Frame: A data packet transmitted in a wireless network. The packet enables a client to reconnect to an access point. The packet is transmitted after a client disconnection or when a client roams from one access point to another.

Virtual Private Network (VPN): A set of technologies and protocols used to establish encrypted tunnels between one or more network nodes.

WiFi Alliance: A nonprofit organization, with more than 200 members, devoted to promoting the use and operation of wireless networks. Products associated by the WiFi Alliance are able to interoperate.

Wireless Computer Network: Any computer network that uses wireless technologies based on the IEEE 802.11x standards to transmit and receive data.

Wireless Sensor Networks

Antonio G. Ruzzelli
University College Dublin, Ireland

Richard Tynan
University College Dublin, Ireland

Michael J. O'Grady
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

INTRODUCTION

The origins of networks of sensors can be traced back to the 1980s when DARPA initiated the distributed sensor networks program. However, recent advances in microprocessor fabrication have led to a dramatic reduction in both the physical size and power consumption of such devices. Battery and sensing technology, as well as communications hardware, have also followed a similar miniaturization trend. The aggregation of these advances has led to the development of networked, millimeter-scale sensing devices capable of complex processing tasks. Collectively these form a wireless sensor network (WSN), thus heralding a new era of ubiquitous sensing technology and applications. Large-scale deployments of these networks have been used in many diverse fields such as wildlife habitat monitoring (Mainwaring, Polastre, Szewczyk, Culler, & Anderson, 2003), traffic monitoring (Coleri, Cheung, & Varaiya,, 2004), and lighting control (Sandhu, Agogino, & Agogino, 2004).

A number of commercial WSN platforms have been launched in recent years. Examples include the Mica family (Hill, 2003), Smart-Mesh (http://www.dust-inc.com), Ember (http://www.ember.com), iBeans (Rhee, Seetharam, Liu, & Wang, 2003), Soapbox from VTT (http://www.vtt.fi/ele/research/tel/projects/soapbox.html), Smart-Its (http://www.smart-its.org), and the Cube sensor platform (O'Flynn et al., 2005). As the miniaturization of the constituent components of a WSN continues unabated, power consumption likewise diminishes, thus the current generation of sensors can function perfectly for years using standard AA batteries (Polastre, Hill, & Culler, 2004). Alternative solutions may not require any batteries; for example iBeans (Rhee et al., 2003) coupled with an energy harvester can operate by scavenging energy from tiny vibrations that occur naturally. Miniaturized solar panels are another possible solution for outdoor operation. Production costs of single nodes are estimated to be less than a dollar, a significant cost reduction over the price of older sensor models, thus paving the way for large-scale WSN deployments, possibly consisting of a number of nodes several orders of magnitude greater than that in ad-hoc networks (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002).

BACKGROUND

The main components of a WSN are gateways and sensor nodes. The sensor nodes can relay their sensed data either directly to the gateway or through each other depending on the scale of the network. In turn the gateway can send commands down to the nodes to, for example, increase their sampling frequency. In some networks, when the gateway is tethered to an adequate power supply, a greater transmission range can be achieved. This gives rise to an asymmetry in the data acquisition and control protocols, where control commands are sent directly to the node but the data sent from the node to the gateway is multi-hopped. Of course multi hopping of the control commands from the gateway can be used also.

Multi-hopping, while useful in extending the reach or scale of a WSN and reducing the overall transmission cost with respect to direct communication, does have its limitations. The cost of transmitting a packet can be greatly increased depending on the distance a node is from its gateway. Secondly, since nodes nearest the base station, that is, one hop away, will not only have to send their data but also that of all other nodes greater than a single hop, there will be a greater demand placed on the power supply of these nodes. This means that, in general, a node lifespan is inversely proportional to the number of hops it is away from the base station. To alleviate this problem, multiple gateways can be used, with the nodes only transmitting data to their local station. A second solution creates a hierarchy of nodes with varying power and transmission capabilities. Higher power

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

nodes can act as gateways to the gateways for the lower powered sensors.

DEPLOYMENT CONSIDERATIONS

There are a number of issues to consider when deploying a WSN. The first and perhaps most significant is the number of nodes required in the deployment. The quantity of nodes required will primarily be governed by the size of the area to be monitored and the frequency of the sampling required. In general, the more nodes there are in a given WSN, the better the quality of data; however there is a corresponding increase in the time required for processing. Given that the choice on node density has been made, another factor still remains—the node sampling frequency. The choice of this value will depend on what aspect of the environment is being monitored and the power resources available on each node. If the sampling frequency is too high, more power will be consumed than is necessary. However, if it is too low, important events could be missed. A useful strategy might be to alter this value opportunistically so as to deliver optimum performance.

Getting the sensed data from the node requires the wireless transmission of a data packet, a process that can consume a significant portion of the available power resources. One transmission can occur per sensed value when a real-time picture of the environment is required. Or multiple readings can be bundled into a single message. Alternatively the node may intelligently decide not to transmit a packet if, for example, no change in the sensed value had occurred. This can dramatically increase the longevity of the node, and if this is a universal policy adopted by all nodes in the WSN, the lifespan of the WSN can be extended.

Due to the battery operation of the nodes, power management is critically important to the health of a WSN. Another approach to performing power conservation is to enable redundant sensors to hibernate. The rationale behind this is that a sensor consumes little or no power while asleep, and so the more nodes that are hibernating, the less power is being consumed collectively by the network. A hibernating node cannot forward any sensed data, effectively reducing the spatial sampling frequency. Therefore, caution must be exercised when selecting which nodes to hibernate.

Two broad approaches exist for selecting nodes for hibernation. The first is based on defining a sensing radius for each individual sensor. An area is covered if all points with the sensed area lie within the sensing range of at least one sensor. When there are points covered by more than one sensor, it may be possible to hibernate redundant sensors without breaking the coverage constraint.

An alternative approach uses the data being received by nodes. It is based on interpolation and assumes that the required node density exists. Given a collection of nodes, it is possible for them to interpolate the sensed medium at a required point, assuming an interpolation function exists. By interpolating at the point of an individual node, an interpolated value can be obtained. By comparing this to the actual sensed value, an interpolation error can be derived. If this error is less than a particular threshold, then this node is deemed redundant and will hibernate for a predefined period of time before rechecking its redundancy.

Another fundamental issue for practical WSNs is that of sensor calibration (Whitehouse & Culler, 2003). When two nodes observe different values in their sensed data, is it because they are seeing different events or because one or both of the sensors has malfunctioned? Of course calibration can be done prior to deployment, but if the malfunction causes its accuracy to degrade over time, then a recalibration must occur on the fly after deployment. This is a significant problem, since the environment in which the nodes are sensing usually cannot be controlled for the calibration to occur. When sampling the environment, it may be a requirement for all the sensors to sample at the same point in time. This requires a clock synchronization technique that will work over the entire network, and this is quite a difficult task to perform on such computationally challenged devices.

Multi-hop routing can introduce a considerable lag between the time a message is sent from the node and the time it is received at its destination. When the destination is a gateway, it will in turn send control commands to the network based on the data it receives. These control commands may also be multi-hopped to their destination. The aggregate delay can be unacceptable and is usually symptomatic of an overburdened gateway. The introduction of an additional gateway that efficiently partitions the network would alleviate this issue.

A final issue with a practical deployment of a WSN is that of programming and debugging the nodes themselves. A node considered to have one of the richest user interfaces consists of three LEDS, allowing eight program execution states to be displayed at any given point in time. To alleviate this, a methodology has been developed to allow the development of applications off the nodes so that accuracy of the approach can be verified (Tynan, Ruzzelli, & O'Hare, 2005).

APPLICATION CONSIDERATIONS

The autonomous nature of the wireless sensor networks makes the technology versatile with respect to applications. Sensors can be effectively deployed for monitoring and detecting of malfunctioning industrial machinery during normal production activity. Moreover, as no infrastructure is needed, their deployment is immediate and highly adaptive to the environment in which they operate. By means of distributed sampling, sensor networks are able to provide a more accurate and in-depth evaluation of the state of the environment at any moment in time.

Sensors can also be programmed to make decisions at a local stage or through a centralized approach. In the first case, nodes are organized in clusters in which a sensor head is elected. Collected data is sent to the sensor head, where it is evaluated before an appropriate decision is made. For instance it can actuate supplementary machinery in cases where production overloading is occurring. In cases where a sensor is not endowed with a decision-making capability, or in particular circumstances where it does not have enough information to make an immediate decision, collected data is sent to the closest gateway for further processing, and the resulting decision is returned to the sensor. In an energy-saving maneuver, data is sent to neighboring nodes that forward it, via a routing procedure, towards the gateway. Such a scenario is an illustration of a centralized approach of decision making.

For more effective management of the production activity, sensors can interact, possibly using either fixed networked or wireless technology, with personal digital assistants (PDAs) or mobile phones. In this way, wireless sensor networks can improve the supervision of an activity and communicate with a user in a more effective and efficient manner. For example, in the case of industrial machine monitoring, should the sensor head sense that a particular machine has excessive vibration or that a temperature exceeds a certain threshold, it may decide to raise an alarm and call a technician for assistance. However, the approach it takes to this may vary. Flagging an alarm in a centralized control system can be implemented quite easily. However, technicians and maintenance staff are generally mobile, as their duties call them to various places in the factory floor. Thus, the sensor head must be able to route a message to them. As technicians are likely to be equipped with PDAs or mobile phones, instant messaging and SMS are two obvious methods of contacting them.

Applications of m-commerce through the use of collaborative wireless sensor networks and PDAs or mobile phones are numerous. In the case of sensors deployed in a shopping center, a shopper with a PDA or smart phone can request a particular product with certain requirements, for example, model, price, and so on. By means of sensor collaboration, products in the shopping center that match the user requirements can be identified. The shopper can then decide whether to buy remotely or go and view the items in question before buying them. Alternatively, a reverse auction could be initiated with the shops in the mall all vying for the shopper's business. While such a scenario illustrates the potential synergic interaction of WSNs and standard mobile devices, a number of issues must be resolved before such a vision can become a reality.

FUTURE TRENDS

Though significant research in WSNs and mobile computing continues, issues concerning the enablement of seamless and transparent interaction between each domain need to be resolved. A number of issues are now identified.

Communication Protocol Issues

In order for a PDA to communicate with a sensor network, it is necessary that both PDAs and WSNs use the same communication protocol. At present, off-the-shelf PDAs have the Bluetooth protocol for short-range communication provided. Unfortunately, studies of the Bluetooth architecture (Leopold, Dydensborg, & Bonnet, 2003) showed the unsuitability of such a protocol for wireless sensor networks. On the other hand, although recent advances propose a vast number of protocols tailored to WSNs, the communication compatibility between the two technologies is still an open issue.

Ontology Issues

Such kinds of issues arise after PDAs and sensors agree which communication protocol to use. In the context of knowledge sharing between PDAs and sensors at the application layer, they should agree with the specification of a conceptualization, also known as an ontology. Although some research proposes the study of semantic techniques for wireless sensor networks (Whitehouse, Liu, & Zhao, 2006), a comprehensive methodology of PDA/sensor interaction is still an open issue to be addressed.

Trust Management Issues

Requests of m-commerce-related information from sensors to PDAs and vice versa raise issues of trust management. In fact, sensors should trust the quality of service offered by the PDA protocol. On the other side, PDAs should trust sensors when, for example, product availability or machinery conditions are sent to a PDA. While the latter case can be considered as an instance of Internet trust management, the former case needs to consider the issue of memory capacity constraints of sensors. Procedures for realizing trust management on individual sensors, for example, through intelligent agent technologies, need further research.

The big “umbrella” of trust management also includes more specific issues of security. In fact, the multi-hop routing of WSNs together with the relatively simple architecture of sensors pose an inherent risk, as an attacker may only need to compromise one device to compromise the security of the entire network. This concern is amplified in applications like m-commerce where private credentials must be fully safely encoded.

CONCLUSION

Significant advances have been made in WSNs over the last decade. Nevertheless, power consumption remains a significant barrier to their widespread deployment. However, significant strides are being made in this area. Though a mature research discipline in its own right, the issues of interaction and interoperability with conventional computing systems, and mobile computing devices in particular, is becoming increasingly important. In extending the reach of computing technologies into what are frequently remote and hostile environments, WSNs will enable an array of new and innovative applications and services for mobile users.

REFERENCES

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: A survey. *Computer Networks Journal, 38*(4), 393-422.

Bychkovskiy, V., Megerian, S., Estrin, D., & Potkonjak. M. (2003). A collaborative approach to in-place sensor calibration. *Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN'03)*, Palo Alto, CA.

Coleri, S., Cheung, Y., & Varaiya, P. (2004). Sensor networks for monitoring traffic. *Proceedings of the Allerton Conference on Communication, Control and Computing,* Illinois.

Hill, J. (2003). *System architecture for wireless sensor networks.* PhD thesis, University of California–Berkeley, USA.

Leopold, M., Dydensborg, M. B., & Bonnet, P. (2003). Bluetooth and sensor networks: A reality check. *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems,* Los Angeles, CA.

Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D., & Anderson, J. (2002). Wireless sensor networks for habitat monitoring. *Proceedings of the ACM International Workshop on Wireless Sensor Networks and Applications,* Atlanta, GA.

O'Flynn, B., Barroso, A., Bellis, S., Benson, J., Roedig, U., Delaney, K., Barton, J., Sreenan, C., & O'Mathuna, C. (2005). The development of a novel miniaturized modular platform for wireless sensor networks. *Proceedings of the IPSN Track on Sensor Platform, Tools and Design Methods for Networked Embedded Systems (IPSN2005/SPOTS2005)*, Los Angeles, CA.

Polastre, J., Hill, J., & Culler, D. (2004). Versatile low power media access for wireless sensor networks. *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems (SenSys '04)* (pp. 95-107). New York.

Rhee, S., Seetharam, D., Liu, S., & Wang, N. (2003). iBean network: An ultra low power wireless sensor network. *Proceedings of UbiComp, the 5th International Conference on Ubiquitous Computing,* Seattle, WA.

Sandhu, J., Agogino, A., & Agogino, A. (2004). Wireless sensor networks for commercial lighting control: Decision making with multi-agent systems. *Proceedings of the AAAI-04 Workshop on Sensor Networks,* San Jose, CA.

Tynan, R., Ruzzelli, A. G., & O'Hare, G. M. P. (2005). A methodology for the development of multi-agent systems on wireless sensor networks. *Proceedings of the 17th International Conference on Software Engineering and Knowledge Engineering (SEKE'05)*, Taiwan.

Whitehouse, K., & and Culler, D. (2003). Macro-calibration in sensor/actuator networks. *Mobile Networks and Applications Journal (MONET),* (Special Issue on Wireless Sensor Networks).

Whitehouse, K., Liu, J., & Zhao, F. (2006). Semantic streams: A framework for composable inference over sensor data. *Proceedings of the European Workshop on Wireless Sensor Networks (EWSN)*, Zurich, Switzerland.

KEY TERMS

Gateway: In general, considered a more powerful node that is used to collect information from the networks and for some architecture to synchronize them. A WSN can have few gateways deployed in the network. Sometimes, they are assumed to be interconnected through alternative wireless technology like WLAN and WiMAX. Gateways are often referred to as sinks or inappropriately called base stations.

MAS: Multi-agent system.

Node Lifespan: Also known as node lifetime, it represents the operational life expectancy of a sensor. Usually, it is calculated for certain network parameters (e.g., network topology, network density) and certain node parameters (e.g., node data rate, duty cycle of a sensor equipped with two AA standard batteries).

Personal Digital Assistant (PDA): Also known as a palmtop.

WSN: Wireless sensor network.

Wireless Technologies for Mobile Computing and Commerce

David Wright
University of Ottawa, Canada

INTRODUCTION

At the time of writing (1Q06) most countries have a small number (2-6) of major cellular operators offering competing 2.5G and 3G cellular services. In addition, there is a much larger number of operators of WiFi networks. In some cases, a major cellular operator, for example, Deutsche Telekomm and British Telecom, also offers a WiFi service. In other cases, WiFi services are provided by a proliferation of smaller network operators, such as restaurants, laundromats, airports, railways, community associations and municipal governments. Many organizations offer WiFi free of charge as a hospitality service, for example, restaurants. Cellular services offer ubiquitous, low data rate communications for mobile computing and commerce, whereas WiFi offers higher data rates, but less ubiquitous coverage, with limitations on mobility due to business as opposed to technology reasons.

Emerging networks for mobile computing and commerce include WiMAX and WiMobile (Wright, 2006), which offer higher data rates, lower costs and city-wide coverage with handoff of calls among multiple base stations. These new technologies may be deployed by the organizations that currently deploy cellular and WiFi networks, and also may give rise to a new group of competitive wireless network operators.

This article identifies the capabilities needed for mobile computing and commerce and assesses their technology and business implications. It identifies developments in the wireless networks that can be used for mobile computing and commerce, together with the services that can be provided over such networks. It provides a business analysis indicating which network operators can profitably deploy new networks, and which network operators need to establish business and technology links with each other so as to better serve their customers. The resulting range of next generation service, technologies and network operators available for mobile computing and commerce is identified.

WIRELESS NETWORK ARCHITECTURES

Figure 1 illustrates the network architectures for WiFi, Cellular, WiMAX and WiMobile, including the radio access network on the left and the wired core network on the right.

The cellular architecture is the most sophisticated in that the core network includes a circuit network (for legacy circuit switched voice calls), a packet network (for data calls) and an IP Multimedia Subsystem, IMS (for migration of all traffic onto the Internet).

These three networks essentially allow the cellular operator to maintain control over all calls to and from the mobile device, and hence derive revenue from them. In particular the IMS network contains servers for establishing voice and video calls over IP, authenticating users, maintaining records of the current location of a mobile user, accounting, and security. Cellular operators are migrating traffic from their circuit and packet networks onto the IMS.

By contrast, WiFi (IEEE, 1999a, 1999b, 1999c, 2003), WiMAX (IEEE, 2006; Ghosh et al., 2005), and WiMobile (IEEE, 2006; Lawton, 2005) are simply radio access technologies and do not specify a core network. They therefore allow more direct access from a mobile device to the Internet. In particular, the WiMobile specification, which is under development at the time of writing, emphasizes that its design is being optimized for operation with IP. This more open access to the Internet allows a mobile user to set up, for instance, a VoIP call using a third party service without the involvement of the wireless network operator. As the user moves from one access point to another, the call can be maintained using Mobile IP, involving servers maintained by the user's ISP, not by the wireless network operator. Mobile IP can operate over diverse wireless access technologies as described by Benzaid et al. (2004).

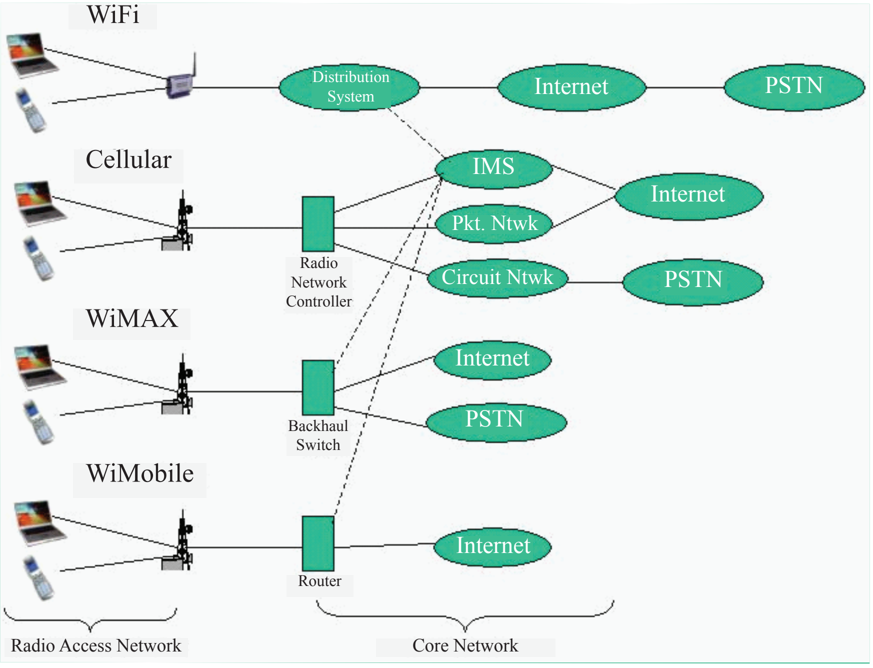
If the operator of a WiFi, WiMAX or WiMobile network wishes to maintain more control over the traffic passing through their network and hence participate more in the revenue generated by that traffic, they can build an IMS network. Alternatively if they already operate a cellular network, they can provide access to their existing IMS network, as shown by the dashed lines in Figure 1.

REQUIREMENTS FOR MOBILE COMPUTING AND COMMERCE

Any wireless transmitter/receiver has a limited range in order to comply with government regulations regarding maximum power output. A mobile user therefore may move out of the

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Wireless network architectures



range of its current wireless access point, and it is necessary to handoff the communication to another access point using either the same or a different wireless technology. Handing off the communication means that the current IP session is maintained, for example, the user continues to browse a Web site as a registered user, a VoIP call is not interrupted, and an enterprise user with a laptop-based secure VPN to an enterprise network continues to use the same VPN. There are four requirements in order to achieve handoff suited to mobile computing and commerce:

1. It must be possible to switch the call from one access point to another
2. If the user is receiving quality of service, QoS, for example, a guaranteed low latency, that QoS is maintained after the handoff, and an acceptable number of packets are lost during handoff.
3. If the access points are operated by different network operators, there must be a business arrangement between them regarding mediation of the billing for the call.
4. The organization deploying the wireless access network must be able to make a profit or to have a business model that focuses on hospitality service.

Requirements 1 and 2 are technology related and are discussed next, followed by the business requirements 3 and 4.

TECHNOLOGY ISSUES

A mobile device that is capable of using multiple wireless access technologies, such as those described above, can continuously scan its radio environment to search for access points that it could potentially use. Some of them may not be available, if, for instance, they are operated by companies with which the user does not have a subscription. In order to choose among the available access points within range the mobile device can apply criteria including: data rate, cost, ability to handoff seamlessly, and QoS; delay (important for voice) and packet loss rate (important for data). For instance, a mobile device with an interactive voice/video call in progress could choose the lowest cost network that provides acceptable delay. A device downloading a large data file could choose the network with the highest data rate given limitations on cost and packet loss rate. Once the network is selected, handoff is initiated.

Handoff among WiFi, WiMAX and WiMobile is handled by IEEE (2006). Handoff between cellular and one of these three technologies is complicated by the need to interwork with the cellular circuit, packet and IMS networks.

- In the case of WiFi, this interworking is provided by a specification from the industry consortium UMA, Unlicensed Mobile Access (2006), which is incorporated as part of the GSM cellular network specifications, release 6.

Table 1. Wireless access network operators revenue sources

Revenue Source	**Service provided by:** **(N) Network Operator** **(C) Content Provider**	**Revenue accrues to:** **(W) Wireless Network Operator** **(3) Third Party Service Provider** **(S) Shared with Content Provider**
Voice/video calls	N	W 3
Audio content	C	S
Video content	C	S
Gateway to PSTN	N	W 3
Geographic info (e.g., travel directions, highway safety)	N C	W S
Location enabled advertising	N C	W S
Location enabled buddy lists	N	W
Multimedia Messaging Service	C	W 3 S
Gaming	C	S
QoS	N	W
VPN	N	W 3

- In the case of WiMAX, similar issues are involved and are being resolved by the WiMAX Forum (2006).
- WiMobile is at an early stage of development and interworking with cellular is not a priority at this stage. A specification may be developed later, or alternatively, WiMobile may differentiate itself from the other technologies by becoming a "native-IP" access mechanism, similar to DSL and cable modem in which customers have direct access to the Internet.

This discussion addresses requirements 1, 2 above. We now move on to requirements 3, 4.

BUSINESS ISSUES

This section presents business strategies for wireless access network operators that take into account sources of revenue related to mobile computing and commerce, plus the need to compete with other technologies and network operators. Earlier work in this area (du Preez & Pistorius, 2003) dates from a time when 3G and wireless data services were emerging technologies. The present section incorporates developments in technology and services to date.

The sources of revenue are given in Table 1 and are classified in two ways:

1. Whether the service is provided by a content provider or a network operator, which may be the wireless access network operator or another network operator. For instance, a VoIP service could be provided by the wireless operator or by a third party such as Vonage. Either way it is provided by a network operator.
2. Who receives the revenue for the service: the wireless access network operator, a third party or a sharing arrangement with a content provider.

It can be seen from Table 1 that there is a large number of mobile computing and commerce services that can be provided by a mix of wireless network operators, content providers and third parties. In addition there are non-revenue generating services such as e-mail and Web browsing. A clear business strategy is needed to operate successfully in competition with the other players. Strategies suited to the different types of wireless network operators are given in Table 2.

Table 2 divides wireless access network operators into three groups: incumbent cellular operators, hospitality providers such as restaurants and municipalities, and new competitors, who are starting operations based on the availability of new technology. The incumbent cellular operators have complex core networks as shown in Figure 1 and incur costs of operating legacy technologies. They seek to deploy all possible wireless technologies in order to accommodate the needs of all customers. By contrast the new competitors seek to reduce their costs by only operating the most recent technologies. Both these groups are operating commercial services and therefore use licensed spectrum so that their customers do not experience interference from other users. The hospitality providers, however, are providing a free service. Their customers accept that the performance may vary according to the demands of other users and therefore the operators reduce their costs by using unlicensed spectrum.

Table 2. Strategy for wireless access network operators

	Cellular Operators	**Hospitality providers**	**Competitive Wireless Network Operators**
Technologies	2.5G, 3G, WiFi, WiMax, WiMobile	WiFi, unlicensed WiMAX	WiMAX, WiMobile
Revenue sources	Generate revenue from the full range of services	Provide Internet access for the full range of services.	Generate revenue from the full range of services
IMS strategy	Lock customers into IMS-based services.	Establish partnerships and interfaces to the IMS of other operators	Build IMS. Establish partnerships and interfaces to the IMS of other operators
Competitive strategy	Buy up competitors.	Avoid competing with other operators by a competitive bid process.	Differentiate from incumbents by offering low cost services, focusing on IP, developing next generation services, for example, presence, location, QoS.

Both the incumbents and the new competitors aim to deliver the full range of services listed in Table 1 to their customers, typically from the IMS, so as to maintain control over the revenue. The hospitality providers, however, are typically providing access only, allowing their customers to get services from any third party they wish, since they do not seek to generate revenue from their networks. For location-based services, the hospitality provider can provide the third party with information about the customer's current location.

The cellular incumbents typically already have an IMS in place and aim to lock customers into service provided by that IMS. The new competitors need to build an IMS and then establish partnerships with other wireless operators so that calls originating on one IMS can be handed off to another operator. These partnerships are also important to the hospitality providers since they typically have no interest in developing their own IMS.

The competitive strategy of incumbent cellular operators towards WiFi operators historically has been to buy them up, and this strategy is also appropriate for WiMAX and WiMobile operators. The strategy of hospitality operators is to avoid competition, and this is particularly important for municipalities, who should not be seen to use tax dollars to compete against private industry. In order to avoid this perception, they can use a competitive bid process allowing any operator the opportunity to bid on the contract to build and operate their network. The strategy of the new competitors is to compete on three fronts. First, they can offer low cost services, since they do not have the cost of operating legacy networks. Second, they can offer a full range of next generation services, such as presence and location-based services, thus positioning themselves as state-of-the-art suppliers. Third, they can sell QoS guarantees to their customers, since new technologies such as WiMAX and WiMobile are particularly suited to providing such guarantees.

CONCLUSION

The enabling technologies for mobile computing and commerce are developing rapidly. New wireless technologies such as WiMAX and WiMobile offer extended coverage and improved QoS compared to WiFi; and higher data rates and lower costs compared to 2.5G and 3G cellular. A wide range of services is available over these technologies including services that generate revenue (a) for the wireless operator, such as location-based services, (b) for a third party, such as VoIP and (c) for a content provider, such as entertainment. Wireless network operators, including incumbent cellular operators, hospitality providers and new competitive wireless network operators, need to develop strategies that allow handoff of calls among the different technologies and operators. Strategies include locking customers into an IMS, interworking with other operators' IMSs, buying out competitors and developing a broad range of state-of-the-art services such as location and presence services.

The mobile computing and commerce user can therefore expect a proliferation of services (Table 1), a number of different network operators (Table 2), an array of different wireless technologies, WiFi, 3G, WiMAX and WiMobile, and a mobile device that can make the best choice among these alternatives at any point in time and space.

REFERENCES

Benzaid, M., Minet, P., Al Agha, Kh., Adjih, C., & Allard. G. (2004). Integration of mobile-IP for universal mobility. *Wireless Networks*, *10*(4), 377-388.

du Preez, G. T., & Pistorius, C. W. I. (2003). Analyzing technological threats and opportunities in wireless data services. *Technological Forecasting and Social Change, 70*(1), 1-20.

Ghosh, A., Wolter, D. R., Andrews, J. G., & Chen, R. (2005, February). Broadband wireless access with WiMax/802.16: Current performance benchmarks and future potential. *IEEE Communications*, *43*(2), 129-136.

IEEE. (1999a). *802.11 Wireless LAN: Medium access control (MAC) and physical layer (PHY) specifications*. New York: IEEE Publications.

IEEE. (1999b). *802.11a high-speed physical layer in the 5 GHz band*. New York: IEEE Publications.

IEEE. (1999c). *802.11b higher-speed physical layer (PHY) extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2003). *802.11g further higher-speed physical layer extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2006a). *802.16e air interface for fixed and mobile broadband wireless access systems: Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands*. New York: IEEE Publications.

IEEE. (2006b). *802.20 mobile broadband wireless access* (In Progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/20

IEEE. (2006c). *802.21 media independent handover services* (In Progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/21/

Lawton, G. (2005). What lies ahead for cellular technology? *IEEE Computer*, *38*(6), 14-17.

UMA. (2006). *Unlicensed mobile access*. Retrieved from http://www.umatechnology.org/specifications/index.htm

WiMAX Forum. (2006). Retrieved March 2006, from www.wimaxforum.org

Wright, D. (2006). Wireless technologies for mobile computing and commerce. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce*. Hershey, PA: Idea Group Reference.

KEY TERMS

IMS, IP Multimedia Subsystem: Part of the wired core network containing servers for establishing voice and video calls over IP, authenticating users, maintaining records of the current location of a mobile user, accounting, and security.

Location-Based Services: Services that take into account the users current geographical location, for example, advertising locally available products and services, providing directions and alerting drivers to traffic congestion and road accidents.

Mobile IP: An Internet standard that allows a mobile user to move from one point of attachment of the network to another while maintaining an existing TCP/IP session. Incoming packets to the user are forwarded to a server in the user's new access IP subnetwork.

Presence: The ability of a user device to specify characteristics, such as whether the user is online, whether the user is willing to receive calls, whether the user is willing to receive calls of a given type (e.g., voice, video, data, MMS) from specified other users and what is the user's current location to a specified degree of accuracy.

Quality of Service (QoS): Features related to a communication, such as delay, variability of delay, bit error rate and packet loss rate. Additional parameters may also be included, for example, peak data rate, average data rate, percentage of time that the service is available, mean time to repair faults and how the customer is compensated if QoS guarantees are not met by a service provider.

WiFi: A commercial implementation of the IEEE 802.11 standard in which the equipment has been certified by the WiFi Alliance, an industry consortium.

WiMAX: A commercial implementation of the IEEE 802.16 standard in which the equipment has been certified by the WiMAX Forum, an industry consortium.

WiMobile: Another name for the IEEE 802.20 standard, which is in course of development at the time of writing (1Q06).

Workflow Management Systems in MANETs

Fabio De Rosa
University of Rome "La Sapienza", Italy

Massimiliano de Leoni
University of Rome "La Sapienza", Italy

Massimo Mecella
University of Rome "La Sapienza", Italy

INTRODUCTION

The widespread availability of network-enabled handheld devices (e.g., PDAs with WiFi) has made *pervasive computing* environment development an emerging reality. Mobile (or multi-hop) Ad-hoc NETworks (MANETs—Agrawal & Zeng, 2003) are mobile device networks communicating via wireless links without relying on an underlying infrastructure. Each device in a MANET acts as an endpoint and as a router forwarding messages to devices within radio range. MANETs are a sound alternative to infrastructure-based networks whenever the infrastructure is lacking or unusable, for example, military applications, disaster/relief, emergency situations, and communication between vehicles.

Generally, the use of a MANET requires a strong collaboration among users/devices constituting the network; more often, that collaboration is translated into a list of activities executed in sequence or concurrently by applications running on mobile devices, thus resulting in *cooperative works*. Therefore, in such situations, MANET users would benefit from software supporting their collaboration. Such a *coordination software* would enable them to execute their *activities* through specific applications—for example, computer-supported cooperative work tools (Grudin, 1994), *workflow management* applications (Leymann & Roller, 2000), and so forth—running on handheld devices, thus enabling cooperative processes to be run. All such applications typically require continuous inter-device connections (e.g., for data/information sharing, activity scheduling and coordination, etc.), but these are not generally guaranteed in MANETs, due to the high mobility of the nodes in the network.

Collaborative Scenarios

Consider a scenario of emergency situations, for example, the case of an aftermath of an archeological disaster. A team is equipped with mobile devices (laptops and PDAs) and sent to the affected area to evaluate the condition of archeological sites and buildings, with the goal of drawing a situation map to schedule rebuilding activities. A typical cooperative process to be enacted by the team would be as shown in Figure 1 (depicted as a UML Activity Diagram—De Rosa, Malizia, & Mecella, 2005):

1. The team leader has previously stored all area details (not included in the process), including a site map, a list of the most important objects located in the site, and previous reports/materials.
2. The team is considered as an overall MANET, in which the team leader's device (requiring the most computing power, therefore usually a laptop) coordinates the other team members' devices, by providing suitable information (e.g., maps, sensitive objects, etc.) and assigning activities/tasks.
3. Team members are equipped with handheld devices (PDAs), which allow them to run some operations but do not have much computing power. Such operations, possibly involving various hardware items (e.g., digital cameras, GPRS connections, computing power for image processing, main storage, etc.) are provided as software services to be coordinated. Team member 1 might fill in some specific questionnaires (after a visual analysis of a building), to be analyzed by the team leader using specific software in order to schedule subsequent activities; team member 3 might take pictures of the damaged buildings, while team member 2 may be responsible for specific processing of previous and recent pictures (e.g., for initial identification of architectural anomalies).

In this case, it might be useful to match new pictures with previously stored images. The device holding the high-resolution camera must therefore be connected to the one containing the stored pictures. But in a situation such as the one shown in Figure 2, the movement of the operator/device equipped with the camera may result in its disconnection from the others.

A pervasive architecture should be able to predict such a situation, alert the coordination layer, and possibly have a

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Cooperative process

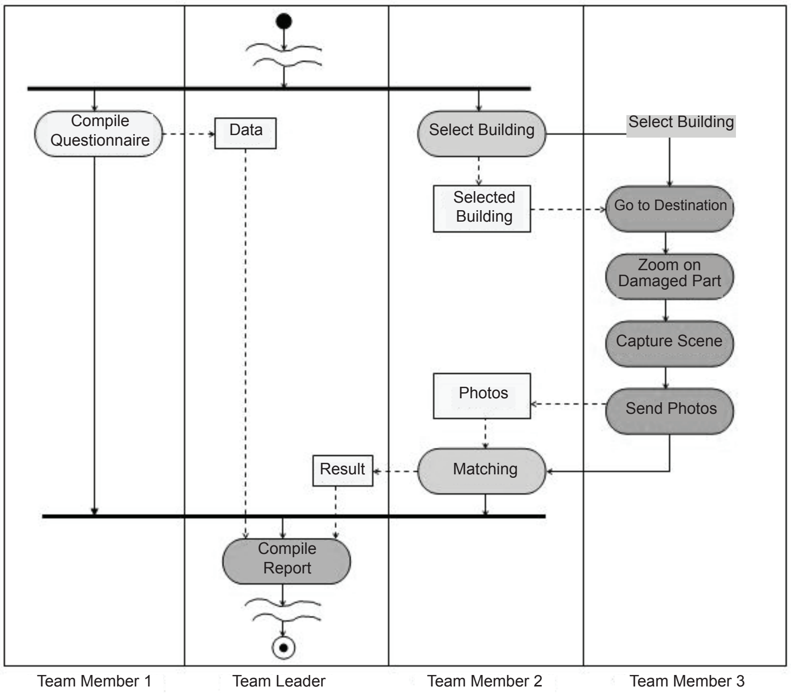
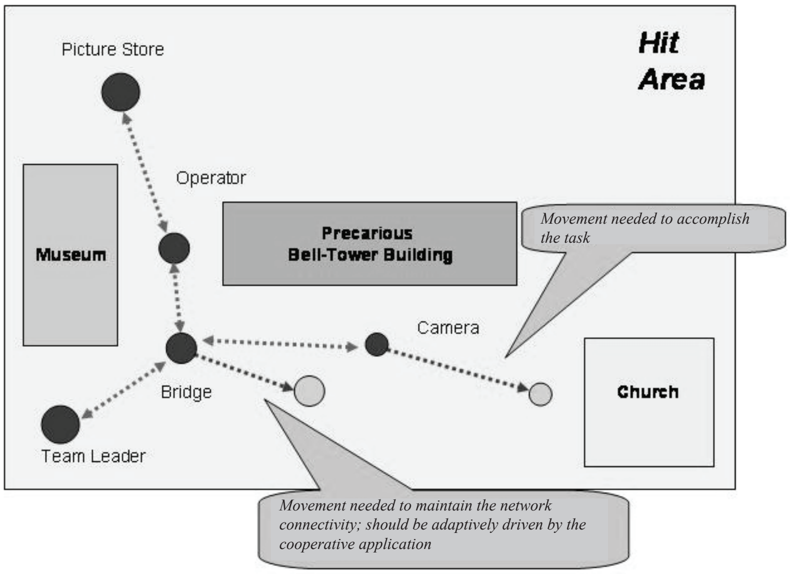

Compile Questionnaire
Data
Select Building
Select Building
Selected Building
Go to Destination
Zoom on Damaged Part
Capture Scene
Photos
Send Photos
Result
Matching
Compile Report
Team Member 1
Team Leader
Team Member 2
Team Member 3

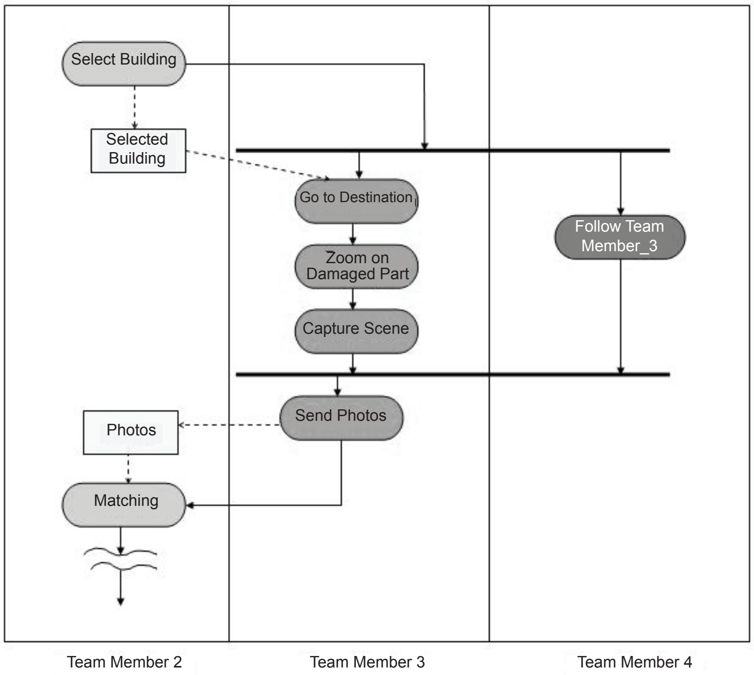
Figure 2. Critical situation and adaptive management


Picture Store
Hit Area
Operator
Museum
Precarious Bell-Tower Building
Movement needed to accomplish the task
Camera
Bridge
Church
Team Leader
Movement needed to maintain the network connectivity; should be adaptively driven by the cooperative application

"*bridging*" device (team member 4's device) to follow the operator/device moving out of range, maintaining the connection, and ensuring a path between devices. In this way the coordination layer schedules the execution of new activities based on the prediction of a disconnection, as shown in Figure 3 (note the new activity for team member 4).

The process's adaptive change is centrally managed by the coordination layer, which has "global" knowledge of the status of all operators/devices and takes into account idle devices, operations that can be safely delayed, and so forth.

Figure 3. Modified process (details)



Some Considerations

Cooperative work may be performed by Workflow Management Systems (WfMSs) (Hollingsworth, 1995; van der Aalst & Van Hee, 2002), a software layer for defining, handling, and carrying out computerized processes (referred to as *workflows*), through programs whose performing order is leaded by a process logic representation.

A workflow is made up of a set of tasks to be performed, a set of state variables handled by tasks, and a set of conditions on such variables which can cause a selection in performing between two or more tasks. Occasionally, some tasks have to be carried out strictly in sequence; other times in a parallel or arbitrary order because there is no causal relationship between them. Actors performing workflow tasks exchange much information needed for their collaboration and coordination. Such an exchange can occur, of course, only if devices/actors are continuously linked to each other.

In a MANET scenario, the "bridging" actions can be seen as supporting activities needed for primary ones which, otherwise, could not be carried out, prejudicing the whole process execution. These support activities are added, when needed, concurrently with other ones. Therefore, a WfMS for MANET has to support dynamic adaptation of processes where the definition of process instances changes on altering of execution context. In MANETs, the altering is caused by disconnection of devices.

The critical issue of current WfMSs is the lack of support of adaptiveness; most WFMSs (both commercial products and academic prototypes) handling adaptiveness assume the presence of an expert who decides which changes have to be applied. In MANETs, the assumption of an expert whose only purpose is disconnection handling is not acceptable, therefore we must investigate how automatically to decide, first, the inadequacy of process instance, and, then which changes on workflow definition must be applied to handle disconnections.

Related and Background Works

In recent years, research in the MANET area has mainly focused on the development of appropriate routing protocols, security and reliability of the communications, methods for energy preservation, and other issues on the lower four ISO/OSI layers (Arbaugh, 2004; Vaidya, 2004). Effective routing in ad-hoc networks is still an actively addressed open problem (Beraldi & Baldoni, 2002; Vaidya, 2004), with some interesting proposals presented in the literature (e.g., dynamic source routing—DSR, ad-hoc on-demand distance vector—AODV routing, zone routing protocol—Z-RP, etc.). Researchers in this area assert that a sound technical basis for MANETs exists and it is thus time to start thinking about how to support applications based on MANETs. In order to enable the development of application layer software (and thus of any information system for MANETs), abstractions on the specific characteristics of the routing algorithms, and more generally, on the services and data provided by the lower network layers, are required. (De Rosa, Di Martino, Paglione, & Mecella, 2003) proposes a network service interface to be used as the basic layer on which to build application software, starting from the analysis and abstraction of current routing protocols.

As far as the issue of adaptive workflow management (Baresi, Casati, Castano, Mirbel, & Pernici, 1999; Voorhoeve & van der Aalst, 1997), this is still an open issue. Relevant work includes the e-Flow system (Casati & Shan, 2001), in which the issue of manually modifying workflow schemas and then automatically migrating active process instances to the new schema is addressed, and AGENTWORK (Müller, Greiner, & Rahm, 2004), which is one of the few examples of a workflow system in which adaption is not manual, but automatic, on the basis of a rule-base approach. But in MANETs the adaption should be carried out in a very frequently changing environment, whereas the previous approaches are targeted to Web-based workflows (indeed workflows composed of different Web services), in which modifications of the schemas are less frequent, but the number of running instances is very high.

THE MOBIDIS PERVASIVE WfMS

Our approach, named MOBIDIS[1], is based on the following specific assumptions:

- Each device includes hardware that lets it know its *communication distance* from the surrounding devices that are within radio range. This is not a very strong assumption, because specific techniques and methods are easily available—for example, TDOA (time difference of arrival), SNR (signal-to-noise ratio), and the Cricket compass (Priyantha et al., 2001).
- No device in the MANET has GPS hardware.
- At start-up, all devices are connected (that is, each device has a path—possibly multi-hop—to any other device). Each device does not have to be within range of any other device; that is, it does not require a *tight* (one-hop) connection. It requires only a *loose* connection, guaranteed by appropriate routing protocols.
- A specific device in the MANET, called the *coordinator;* centrally predicts disconnections and manages them by rewriting the workflow schema and (if it is required) by reassigning the process tasks among the participants.

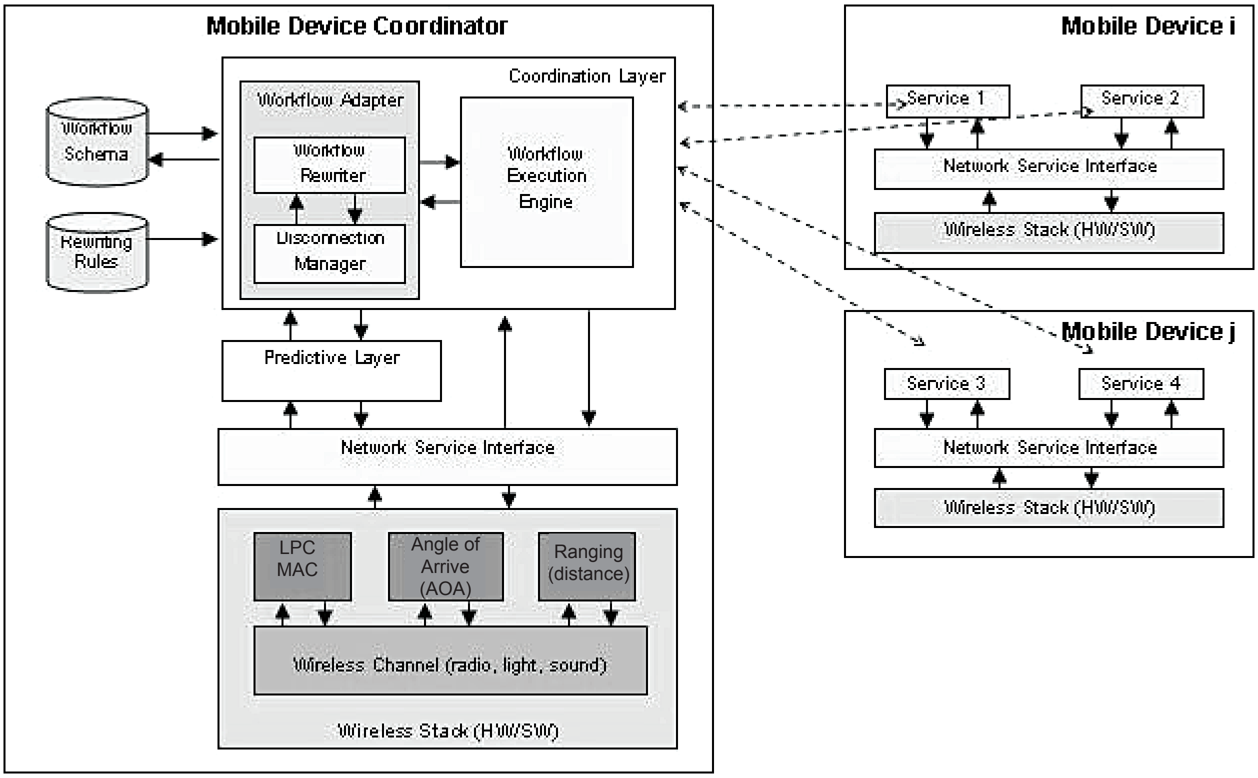
The proposed approach combines local connection management among devices with global management of both network topology and task assignment. Local connection management consists of monitoring and checking one-hop communications between a device and its neighbors. It is realized as special services running on handheld devices that implement techniques for estimating and calculating distances and relative positions (angle and direction of arrival) between a specific device and its direct neighbors. Global management maintains a consistent state of the network and of each peer in the network. It manages the network topology (and its predicted next states) and the tasks each peer is in charge of, as well as services that peers offer (that is, it provides a service registry). On the basis of that information, the coordinator applies algorithms for choosing a bridge and/or executes workflow task reassignment when needed.

Figure 4 shows the proposed architecture to support disconnections and to enact cooperative work in MANETs: each device has a *wireless stack* consisting of a wireless network interface (the *wireless channel* and *LPC MAC* modules) and the hardware for calculating distances from neighbors (*angle of arrive* and *ranging* modules).

On top, a *network service interface* offers to upper layers the basic services for sending and receiving messages (through multi-hop paths) to and from other devices, by abstracting over the specific routing protocols. Offered services (i.e., specific applications supporting tasks of the devices' human users) are accessible to other devices and can be coordinated and composed cooperatively. Some of these services are applications that do not require human intervention (for example, an image-processing utility). Others act as proxies for humans (for example, the service for instructing human users to follow a peer is a simple GUI that alerts the user by displaying a pop-up window on his or her device and emitting a signal).

In contrast, the coordinator device presents the *predictive layer* on top of the network service interface, signaling any probable disconnection to the upper c*oordination layer.* The *predictive layer* implements a probabilistic technique (De Rosa et al., 2005) which can predict if all devices will still be connected in the next instant. At a given time instant t_i in which all devices are connected, the coordinator device collects all device distance information and builds a next-connection-graph—that is, the most likely graph at the next time instant t_{i+1}, in which the predicted connected and disconnected devices are highlighted. In the interval $[t_i, t_{i+1}]$, the coordinator layer enacts the appropriate actions to enable all devices to be still connected at t_{i+1}. In predicting at t_i the next-connection-graph, the technique considers not only the current situation, but also recent situations and predictions (i.e., at t_{i-1}, t_{i-2}, etc.), specifically considering distances calculated in the recent past. Thus, although the pervasive architecture guarantees the connection of all devices, MANET's evolution is considered as it would be in a "free" scenario (i.e., without remedial actions by the coordination layer) when predicting the future situation. The reasonable assumption is that if two devices have the

Figure 4. MOBIDIS architecture


Mobile Device Coordinator
Coordination Layer
Workflow Adapter
Workflow Schema
Workflow Rewriter
Workflow Execution Engine
Rewriting Rules
Disconnection Manager
Predictive Layer
Network Service Interface
LPC MAC
Angle of Arrive (AOA)
Ranging (distance)
Wireless Channel (radio, light, sound)
Wireless Stack (HW/SW)
Mobile Device i
Service 1
Service 2
Network Service Interface
Wireless Stack (HW/SW)
Mobile Device j
Service 3
Service 4
Network Service Interface
Wireless Stack (HW/SW)

tendency to go out of radio range if left "free," and are thus connected through the coordinator's remedial actions, then this influences the subsequent connection probability.

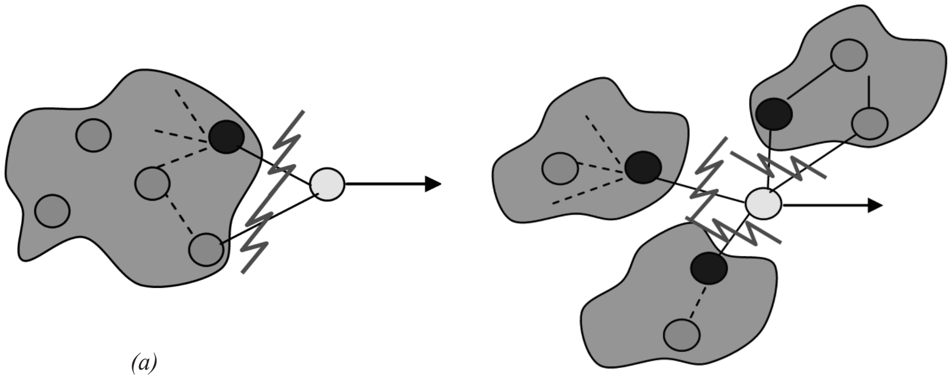
The coordination layer manages situations when a peer is going to disconnect, by applying algorithms for choosing a *bridge,* and by executing workflow schema restructuring and workflow task reassignment when needed (e.g., it assigns the activity "follow peer *x*" to the selected bridge). The algorithm for choosing a bridge selects the best one by using the following criteria:

1. The algorithm chooses as bridge the neighbors without performing tasks. Indeed, if the selected one was a neighbor carrying out a task, its performing activity should be rolled-back with a decrease of productivity.
2. If each neighbor is carrying out a task, then the lowest priority task is chosen. By rolling-back task with the lowest priority, inefficiency should be bounded.
3. If two or more actors perform a task with the same lowest priority, the one with the smallest number of neighbors is preferred. The bridge role likely leads to movement of the node and this might cause new disconnections. By selecting a node with the lowest number of neighbors, the probability of new disconnection is minimized.
4. In the end, the nearest neighbor is preferred.

When the bridge node is selected, the process instance is modified by adding the supporting activity "follow **x**" concurrently with the activity of node **x** going to disconnect. Thus the supporting activities have the same priority of supported ones. This is the situation depicted in Figure 5a in which the bridge node is blue colored. In some cases, the node going to disconnect can produce more than two components (Figure 5b) in the network. In this situation, a bridge is needed for each partition the disconnection is going to create. For each partition, the bridge is selected by applying an algorithm instance over the subset of node belonging to such a partition.

An important issue of the bridge algorithm is the priority among activities. The priority must reflect the process structure (causal dependency): the purpose is to assign higher priority to those tasks whose executions generate the achievement of *enabled* state for a greater number of tasks. An *enabled* task is a not yet *running* task that is ready to be assigned to an actor, as all the preceding tasks have been terminated.

Figure 5. Creation of two (a) or more (b) partitions



(a)

(b)

The algorithm for computing priorities is based on an *n*-ary tree that can be built iteratively. A well-structured process can be shown as a composition of many sub-processes decomposable in other smaller sub-processes and so on up to elementary processes—that is, activities. Each node of the tree is a process (elementary or not) whose children are nodes representing the sub-processes it can be decomposed in. The weight of leaves is initially posed to 1; the weight of internal nodes is obtained by combination of weights of children nodes, according to the way the process is decomposed.

Therefore the restructuring of the process is reduced to a transformation of such an *n*-ary tree.

FUTURE TRENDS

In the context of some Italian and international research projects, we are going to validate our approach in real scenarios. Preliminary experiments on synthetic data shows the feasibility of the approach (De Rosa et al., 2005), but an extensive real validation is needed.

Moreover, we will also address the issue of the approach's fault tolerance. Our approach currently does not cope with sudden downs of devices, which might be frequent in emergency scenarios and are critical if they affect the coordinator node. We also plan to evolve the coordination layer from a centralized to a distributed one (i.e., having a subset of devices act as coordinators). At the moment, the centralized architecture might be a bottleneck, but the current dimensions of a typical MANET for the considered scenarios (tens of devices) do not pose critical scalability issues.

In the future, due to the wide diffusion of mobile devices, applications on MANETs will become more and more interesting. As we have shown, such applications should take into account disconnection anomalies. To date, we have investigated how classical WfMS concepts should be evolved and adapted in order to cope with MANET scenarios, but surely similar issues will also be raised for many other classical technologies, when applied to MANETs.

REFERENCES

Agrawal, D. P., & Zeng, Q. A. (2003). *Introduction to wireless and mobile systems.* Thomson Brooks/Cole.

Arbaugh, W. A. (Ed.). (2004). Making wireless work. *IEEE Security & Privacy, 2*(3), 7-96.

Baresi, L., Casati, F., Castano, S., Mirbel, I., & Pernici, B. (1999). WIDE workflow development methodology. *Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration* (pp. 19-28).

Beraldi, R., & Baldoni, R. (2002). Unicast routing techniques for mobile ad hoc networks. *The handbook of mobile ad hoc networks.* CRC Press.

Casati, F., & Shan, M. C. (2001). Dynamic and adaptive composition of e-services. *Information Systems, 6*(3), 143-163.

De Rosa, F., Di Martino, V., Paglione, L., & Mecella, M. (2003). Mobile adaptive information systems on MANET: What we need as basic layer? *Proceedings of the 1st IEEE Workshop on Multichannel and Mobile Information Systems (MMIS'03),* Rome (pp. 260-267). IEEE CS Press.

De Rosa, F., Malizia, A., & Mecella, M. (2005). Disconnection prediction in mobile ad hoc networks for supporting cooperative work. *IEEE Pervasive Computing, 4*(3), 62-70.

Grudin, J. (1994). Computer-supported cooperative work: History and focus. *IEEE Computer, 27*(5), 19-26.

Hollingsworth, D. (1995, January). *The workflow reference model.* Workflow Management Coalition.

Leymann, F., & Roller, D. (2000). *Production workflow: Concepts and techniques.* Englewood Cliffs, NJ: Prentice Hall.

Müller, R., Greiner, U., & Rahm, E. (2004). AgentWork: A workflow-system supporting rule-based workflow adaptation. *Data and Knowledge Engineering, 51*(2), 223-256.

Priyantha, N. B., Miu, A., Balakrishnan, H., & Teller, S. (2001). The Cricket compass for context-aware mobile applications. *Proceedings of the 7th Annual International Conference Mobile Computing and Networking (MobiCom01)* (pp. 1-14). Boston: ACM Press.

Vaidya, N. H. (2004). *Mobile ad hoc networks: Routing, MAC and transport issues.* Tutorial on Mobile Ad Hoc Networks, University of Illinois at Urbana-Champaign, USA. Retrieved from http://www.crhc.uiuc.edu/nhv

van der Aalst, W., & Van Hee, K. (2002). *Workflow management: Models, methods and systems.* Cambridge, MA: The MIT Press.

Voorhoeve, M., & van der Aalst, W. (1997). Ad-hoc workflow: Problems and solutions. *Proceedings of the 8th DEXA Conference.*

KEY TERMS

Adaptive WfMS: A system that provides process support like normal WfMS, but in such a way that it is able to deal with certain changes in execution context (like node disconnections).

Bridge: A MANET device in a cooperative group following another one *d* in order to avoid *d* going out from the network, by preserving a multi-hop path from every other node to *d*.

Computer-Supported Cooperative Work (CSCW): A collective name for the methods and techniques of a system that support the cooperative performance of work.

MANET (Mobile (or Multi-hop) Ad-hoc NETwork): Mobile network where devices communicate one with another via wireless links without relying on an underlying infrastructure, like an access point.

Signal-to-Noise Ratio (SNR): Difference in decibels between the strength of a signal emitted/received from/to a device and the noise emitted/received from/to the same device. In a radio link, the strength of a signal received decays and the noise grows exponentially with the distance from transmitter (so SNR decays). It is possible to deduce distance between a pair of devices by analyzing SNR of receiving signals.

Workflow Management System (WfMS): A system that completely defines, manages, and executes processes through the execution of software whose order of execution is driven by a computer representation of the business process logic.

ENDNOTE

[1] MOBIDIS: MOBile at Dipartimento di Informatica e Sistemistica (DIS), of the University of Rome "La Sapienza", Italy.

XML-Based Languages for Multimodality in Mobile Environments

Danilo Avola
Istituto di Ricerche sulla Popolazione e le Politiche Sociali, Italy

Maria Chiara Caschera
Istituto di Ricerche sulla Popolazione e le Politiche Sociali, Italy

Fernando Ferri
Istituto di Ricerche sulla Popolazione e le Politiche Sociali, Italy

Patrizia Grifoni
Istituto di Ricerche sulla Popolazione e le Politiche Sociali, Italy

INTRODUCTION

The development of multimodal tools and mobile devices in particular is producing great interest, especially for accessing Web information, performing transactions, and use of services in general. This article considers the different markup languages proposed by the working groups of W3C (World Wide Web Consortium) to manage multimodal interaction and perspectives of multimodal applications and services.

The trend toward the convergence of various methodologies and technologies has developed new devices providing complex services, contributing to the sharing of experiences, and promoting the inclusion of people as community members (Paternò, 2004). This trend is based on the development of mobile devices and their usability, accessibility, portability, and versatility (Kvale, Warakagoda, & Knudsen, 2003).

The usefulness and usability of services, and the ability to access them and information, are the basic elements in the diffusion of Web systems and development of Web multimodal languages. The diffusion and implementation of multimodal services is supported by the activities of the World Wide Web Consortium, aimed at extending interaction modes for different devices and particularly devoted to solving various problems connected with: (1) multimodal Web interaction through the different devices, and (2) practice Web navigation from different devices.

Some W3C working groups focus their activities on issues such as independence from devices, multimodal Web access, and types of contents for multimodal messaging. These specifications allow rich multimodal contents to be transmitted, and are based on the power and extensibility of XML (eXtensible Markup Language) (Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 2004).

XML is highly important in a mobile application environment, as many applications have to manage multimedia contents and need dedicated tools for this. SMIL (Synchronized Multimedia Integration Language) (Solon, McKevitt, & Curran, 2004) was proposed to achieve this goal.

In the early years W3C-MMI (W3C—MultiModal Interaction) focused on multimodal interaction modes such as speech and pen interaction, and providing users with W3C technologies.

W3C develops these technologies by orienting individual interaction modes in order to create mixed-namespace XML documents, such as SVG (scalable vector graphics) (Chatty, Lemort, Sire, & Vinot, 2005) and XHTML (eXtensible HyperText Markup Language) (Musciano & Kennedy, 2003) for visual interaction, and VoiceXML (Voice Extensible Markup Language) (Lucas, 2000) for voice interaction.

However, many other XML-derived languages have helped in the development of mobile services.

The next target is the consideration of the mobile network as an extension of the global Internet network. This article explains the importance of XML and its dialects in a mobile application environment to enable their use by the "various applications/services" (today available on the Web). In fact, different dialects may be needed for different mobile devices depending on their characteristics (such as memory, CPU speed, integrated software engine, etc.). For example, two SVG profiles are defined for cellular phones and PDAs (personal digital assistants): SVG Tiny (SVGT) is suitable for the next generation of cellular phones especially, while SVG Basic (SVGB) is aimed at high-tech devices such as PDAs or smart phones (Andersson et al., 2003).

The pervasive use of mobile devices will be the target for the near future (Branco, 2001), given the trend towards considering the mobile network as an extension of the Internet global network. This scenario promotes the development of new dialects for multimodal interaction through mobile devices. The dialects developed for speech, sketch, and visual

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

interaction are discussed next. An area for future development might focus on interaction through gestures.

XML (eXtensible Markup Language) is a simple, flexible, and powerful markup language, based on text format that allows the development of a potentially unlimited number of innovative multimodal services and applications. It was derived from the more complex, complete SGML (Standard Generalized Markup Language, ISO 8879) (Chamberlin & Goldfarb, 1987), designed for more general purposes. However, XML language is easier to manage, and is genuinely Web oriented and mobile oriented. In other words, XML language is an optimal subset of SGML, constructed in consideration of the possible Web services and applications.

XML can be used to develop several languages taking the specific working context into account. It also plays an important role in the exchange of a wide variety of data, making them available and accessible by Web using computers and mobile devices.

BACKGROUND

The World Wide Web is undergoing continuous development. This has enabled a great expansion in a wide variety of applications and services, covering every human activity. In addition, advanced technology mobile systems are becoming ever more complete, complex devices, which can offer a broad range of Web applications and services originally conceived for personal computers. These two factors explain the need to introduce various multimodal systems (services and applications oriented) to interact with the Web using mobile devices.

In this context, the use of XML-based technology to create powerful, multi-purpose system interfaces is a winning choice. Meta-language XML allows ad hoc language solutions to be developed according to the specific argument.

A multitude of XML "dialects" for multimodal solutions have already been developed, while any others are under current or future development. An exhaustive point of view on XML-based languages is thus not a simple matter. This section provides a panorama of the XML-based languages, considering the different interaction modes, multimediality, and multimodality features.

SMIL (Solon et al., 2004) is a basic, developing technology that allows several multimodal environments to be implemented. It is not an "out and out" multimodal language. As in many multimodal environments, it is necessary to interact with several kinds of multimedia content; SMIL works in the background with many different multimodal applications and services.

SMIL enables interactive audiovisual presentations to be easily produced. It is typically used to choreograph complex multimedia presentations, where audio and video streaming, images, text, graphics, and other media types are combined in real time. In other words, SMIL makes it possible to manage the temporal and spatial constraints of multimedia presentations. The current SMIL conception offers modules for animation, content control, layout, linking, media objects, meta information, timing and synchronization, and transition effects. This modular approach allows reuse of SMIL syntax and semantics in other XML-based languages, especially those used for timing and synchronization. These fundamental features play a leading role in multimodal user interaction. SMIL is an easy-to-learn, HTML-like language and the World Wide Web Consortium (W3C) recommendation to achieve synchronized multimedia. In this context, it simplifies the creation of time-based multimodal interfaces with a high portability factor. SMIL is also exploited to aid the construction of powerful mobile-oriented multimodal applications and services.

A classic example of SMIL use is to enable authors to specify and control the precise time a sentence is spoken and make it coincide with the display of a given image. This simple "technical pattern" is at the base of several multimodal general systems.

To consider the different languages and problems connected with multimodality, we must take into account the different interaction approaches (visual, voice, etc.). For a visual approach, we must consider SVG. The SVG language (Chatty et al., 2005) is another important basic technology that works in background mode to resolve different types of problems in the multimodal interfaces. It describes two-dimensional graphics and graphical applications in XML, providing facilities for document structuring, shape definition, painting, clipping and masking, compositing, text manipulation, styling, linking, scripting, animation, interactivity, integration of multimedia content, alpha masks, filter effects, and template objects. It supports object zooming, interaction and manipulation, and scene annotation, among others. SVG element features can be static or dynamic, and each complex element can be used in interactive mode. A strong point of this language is the ability to interact with script languages (such as javaScript, ECMAScript, JScript, etc.), which provide complete access to all elements, attributes, and properties necessary to develop powerful, complete 2D graphical representations. However, this latter point is not a binding factor. In fact, for example, an animation can be defined and triggered either declaratively (i.e., by embedding SVG animation elements in SVG content) or via scripting.

Because of SVG's ability to produce high-quality rich graphical displays, enable the development of highly interactive user interfaces, and manipulate the contents and structure of an SVG document, it is very well suited for the development of interactive multimodal applications and services.

In fact, it is Web accessible by both personal computer and innovative mobile-oriented systems, and needs sophisticated multimodal environments to satisfy the user's requests. SVG recently became a standard on the Web and is now also becoming a standard for mobile devices. However, different technologies need different hardware supports, as each mobile device has its own characteristics (memory, CPU speed, integrated software engine, etc.). Two different profiles are defined for different device families. The first, low-level profile, SVG Tiny (SVGT), is especially suitable for next-generation cell phones, while the second, SVG Basic (SVGB), is aimed at high-tech devices (such as PDAs or smart phones).

Another markup language, which always works in background mode, is XForms (Extendible Forms) (Bals, 2005). This represents the next generation of forms for the Web and is used for mobile systems. It permits describing general user interfaces, such as Web forms, interactive GUI (Graphical User Interface), special windows, and so forth. It is well structured and can also be used in a standalone manner to describe more complex user interfaces.

The real innovation lies in splitting traditional XHTML forms into three parts: model, instance data, and user interface presentation. The construction of a graphical user interface is the most classic example of a multimodal expression.

MPML (Multimodal Presentation Markup Language) (Zong, Dohi, & Ishizuka, 2000) is another important language developed to enable the description of complex multimodal presentations based on character agents.

This markup language allows a new style of effective presentation of information and a new approach to the production of multimodal information content. These different multimodal scenarios use interactive lifelike agents with expressive capability (e.g., verbal conversations, facial expressions, body gesture explications, behavioral model applications, etc.) that guide users through the multimedia informative content.

MPML language is structured on several defined features, which enhance the capability and potentiality of the multimodal applications and services created with it. Most importantly:

- **It is Independent of the Character Agent System:** It is designed such that authors can write multimodal presentation content independent of specific character agents.
- **It is Easy to Describe:** It is a simple and suitably designed XML-based elaboration designed for multimodal complete presentation only, providing a minimal set of tags to control the presentation and also allowing complete graphical representations.
- **It Supports Interactive Presentation:** It supports a set of mechanisms to guide the presentation interactively.
- **It has Character Control:** Character agents can guide the presentation.
- **It Supports Media Synchronization:** It is in accordance with SMIL, thus enabling the combination and synchronization of various types of multimedia elements such as images, text, animations, audio/video streams, and so forth.

The huge advantage of this language is that authors can provide various features on a single character agent to enable it to adapt to new situations and interactions.

MPML is easy to extend, enabling the expansion of normal application fields and design of multimodal functionality to more complex contexts. For example MPML-VR (Multimodal Presentation Markup Language for Virtual Reality) (Okazaki, Aya, Saeyor, & Ishizuka, 2002), which facilitates multimodal presentation in a 3D virtual space supported from an agent system in VRML (Virtual Reality Modeling Language), was designed by extending MPML. Another versatile extension is MPML-FLASH (Multimodal Presentation Markup Language with Character Agent Control in Flash Medium). Several attempts are ongoing to further widen fields for application of MPML technology.

Another important multimodal XML-based language is EMMA (Extensible MultiModal Annotation Markup Language) (Marrin, Myers, & Aktihanoglu, 2005). Its creators (W3C Multimodal Interaction Working Group) aimed to develop specifications to enable access to the Web (or mobile environment) through multimodal interaction. EMMA is intended for use in systems that provide semantic interpretations for a variety of inputs, including but not necessarily limited to speech, natural language text, GUI, and ink input. Classic examples include prearranged keystroke sequence, speech recognition commands, touch-screen recognition commands, handwriting recognition, semantic vocal streaming understanding, semantic gesture streaming understanding, natural language understanding, interactive ink pens, pointing input tools, and so forth. There is a potentially unlimited possibility to induce numerous input signals from different input devices.

EMMA is used primarily as a standard format for data interchange among the components of a multimodal system. It is automatically generated by interpretation components to represent the semantics of user inputs, not directly authored by developers.

EMMA turns out to be useful in several contexts. It is independent from the specific application field and can be adapted to heterogeneous input types. Different complex devices have their own interaction methods, which can be separately managed to best interact with the user input, and the language provides a set of elements and attributes that focus on accurately representing annotations on the input interpretations. In other words the language focuses on multimodal semantic interpretation of the events that drive user interaction during the composition of one or more inputs.

In general an EMMA document can be considered as consisting of three parts: instance data, data model, and metadata. Instance data is an application-specific markup corresponding to input information meaningful to the consumer of an EMMA document. Instances are built by input processors at runtime. Given that utterances may be ambiguous with respect to input values, an EMMA document may hold more than one instance. The data model imposes constraints on the structure and content of an instance. Metadata represent annotations associated with the data contained in the instance. Annotation values are added by input processors at runtime.

EMMA has a large potential for integrating different devices in several environments. Furthermore, it will be able to make the Web fully accessible for future applications. There should be no difference in interacting with a user via phone DTMF (dual tone multiple frequency) tones, PDA ink pens, or even voice browsers for users with disabilities. This is an impressive goal, and the MIF (multimodal interaction framework, a complex system to support multimodal development of applications and services) is a next step in achieving it.

Another multimodal markup XML-based language is InkML (Ink Markup Language) (Mohamed, Belenkaia, & Ottmann, 2004). It is utilized to build the data format used to represent ink entered with an electronic pen or stylus in a multimodal system. Its main purpose is to transfer digital ink data between devices and software components, and store hand-input traces for handwriting recognition, signature verification, and gesture interpretation.

An increasing number of electronic devices with pen interfaces are now available. These are widely used for various applications and services, such as information, interactive services, semantic interpretation of gestures, multimodal "touched" interaction, and so on. It is also necessary to take into account that handwriting is a very familiar input mode for most users (with or without disabilities), who will thus tend to use it for input and control when available.

Digital ink information is a delicate, complex, and difficult format to manage. Several studies have attempted its comprehensive and productive format standardization. These elements have caused a restricted growth of key technical factors such as digital ink capture, processing, and transmission of digital ink data across heterogeneous devices; communications among different ink management software programs; interactions between ink software applications and hardware devices; and so forth. InkML is also designed to resolve these problems. In fact, its important feature is to provide a simple, platform-independent data format to promote the interchange of digital ink among different kinds of software and hardware applications and services.

An interesting peculiarity of this technology relates to its ability to capture and express the user's dynamic behavior and relative semantics—that is, the dynamics of the user's behavior during the interaction with, for example, the electronic pen on the tablet (to interpret hand-speed, break point, acceleration paths, etc.). To date, only a low information level has been used to interact with the user. It is to be hoped that future studies will examine this scenario. Clearly, this markup language is also useful in both a Web environment and mobile devices. It is becoming a useful standard (W3C recommendation) to tackle entire problematic situations on ink contexts.

The VoiceXML (Lucas, 2000) language provides speech recognition and speech synthesis capabilities for the scripting of voice dialogs and enables integration with other processes through events and IP connectivity. In voice processing, it offers control of voice recognition capabilities through grammars, speech synthesis control, and some basic telephony control. It provides several voice response systems, such as recognition of spoken input and DTMF input. It also allows registration of spoken input and controlling dialogue flow, and can be used to transfer and disconnect telephone calls. VoiceXML also supports the output of synthesized speech and audio files.

This language supports audio file formats, speech grammar formats, and URI schemas.

In order to permit speech applications, VoiceXML is based on a grammar format, called speech recognition grammar specification (SRGS) (Hunt & McGlashan, 2003). SRGS is used by speech recognizers and allows developers to specify the words and word patterns to be listened for by the speech recognizer.

VoiceXML provides semantic interpretations from grammars and makes this information available to the application. It also permits a separation of service logic from interaction behavior.

In IP connectivity VoiceXML allows precise identification of which data to submit to the server, and which HTTP method, GET or POST, to use in the submission.

SSML (Speech Synthesis Markup Language) (Burnett, Walker, & Hunt, 2004) is an XML-based language that presents elements for controlling the pronunciation, tone, inflection, and other characteristics of spoken words. It captures text speed, volume, inflection, and prosody in order to convert it to acoustic speech through a text-to-speech (TTS) synthesis engine. SSML's main features are interoperability—or interaction with other markup languages—and consistency; in fact, it provides control of voice input across platforms and across speech synthesis implementations.

SALT (speech application language tags) (Cisco Systems, Comverse, Intel Corporation, Microsoft Corporation, & Philips Electronics, 2002) is an extension of HTML and other markup languages, integrating speech and telephony interfaces with Web applications and services. In multimodal applications it supports speech input and output in visual pages. It is used where a speech interface is available in addition to the visual interface.

MAIN FOCUS OF THE ARTICLE

The increasing diffusion and utility of the global network has enabled the spread of applications and services supporting every human activity. These involve both scientific domains, such as online technical services, and typical entertainment domains.

The widespread use of mobile devices has improved the diffusion of multifunctional complex tools satisfying sophisticated user applications and increases the spread of innovative services.

In mobile applications, the "audio/video Multilanguage dictionary" enables the user to immediately discover the meaning of idiomatic expressions and difficult words. The multimedia stream allows the meaning of "sentences" in real contexts to be understood, with practical examples and situations.

In this section we discuss some of the most well-known and useful XML-derived languages used to develop mobile services, and give a table that describes these languages according to their modes, functions, and applications/services.

SMIL is the W3C recommendation to synchronize multimedia on the Web. This XML-based language is used to write interactive multimedia presentations, enabling management of their temporal and spatial constraints. In mobile phones, it enables lightweight multimedia functionality and integrates timing into profiles such as WAP forum's WML language and XHTML Basic, while MMS is a subset of SMIL for mobile telephone multimedia messaging.

The W3C group's Voice Browser and Multimodal Interaction are working to standardize XML languages. VoiceXML is used to create voice user interfaces with automatic speech recognition (ASR) and text-to-speech synthesis, audio dialogues that feature synthesized speech; digitized audio; recognition of spoken and DTMF key input; recording of spoken input, telephony, and mixed-initiative conversations; interactive voice response (IVR); and so forth. Many applications and mobile services exploit the potential of VoiceXML, and there are numerous consolidated contributions in this field. In fact, VoiceXML and SALT are valid instruments to support the integration of visual representation enriched systems and visual browsers, with instruments such as XHTML, cascading style sheet (CSS), SMIL, and SVG.

Similar areas are handled by SALT. This language is being designed to "extend existing markup languages such as HTML, XHTML, and XML." An important difference between SALT and VoiceXML is the overall approach used to develop applications. Whereas VoiceXML is essentially declarative, using its extensive set of tags, SALT is very procedural and script oriented, with a very small set of core tags. The factor connecting them is that multimodal access enables users to interact with an application in a variety of different ways. Obviously in this case too, there are a number of practical demonstrations.

New trends are developing to allow intermodal interaction by instruments such as InkML, which captures pen movements and enables data exchange by "digital ink." This mark-up language has an XML data format to describe digital ink data from the pen or stylus in a multimodal system.

A pen-based interface captures digital ink and pen movements by a transducer. The digital ink can be analyzed by recognition software to convert pen input into defined computer actions. Alternatively, it can be stored in ink documents, messages, or notes for later retrieval or exchange through telecommunications.

EMMA is used to implement semantic interpretations of a great variety of inputs, such as voice, text languages, and digital ink. The Speech Services Control (SpeechSC) working group of the Internet Engineering Task Force (IETF) develops protocols to support distributed speech recognition and synthesis and speaker verification services, and expects to take advantage of W3C's work on speech recognition grammar specification (SRGS), SSML, and semantic interpretation (SI).

Table 1 summarizes the main features of each multimodal XML-based language. The table includes modalities that

Table 1. Main features of multimodal XML-based languages

XML-Based Dialects	Modalities	Functions	Application and Services
SMIL	Audio, Video, Images, Text, Graphics	Synchronization, Real-time combining, Managing temporal and spatial constraints, Streaming, Timing	Animation, Content control, Layout, Linking, Media objects, Meta information, Transition effects, Web-oriented multimodal application
SVG *(SVG Tiny—SVGT—for next-generation cell phones, SVG Basic—SVGB—for high-tech devices)*	Graphics	Interactivity, Integration of multimedia content, Two-dimensional graphics, Graphics applications in XML	Document structuring, Definition of shape, Painting, Clipping, Masking, Compositing, Text manipulation, Styling, Linking, Animation, Alpha mask, Filter effects, Zooming, Scene annotation, Object manipulation and interaction
MPML *(MPML-VR, MPML-FLASH)*	Integration complex multimedia, Streaming management	Description of complex multimodal presentations based on character agents, Media synchronization	Web and mobile environment applications, Writing multimodal presentation content independent of specific character agents, Providing a minimal set of tags to control presentation, Interactive presentation guidance

Table 1. continued

EMMA	Speech, Natural language, Text, GUI, Ink input	Semantic interpretations for a variety of inputs	Prearranged keystroke sequence, Speech recognition commands, Touch-screen recognition commands, Semantic vocal streaming comprehension, Semantic gesture streaming comprehension, Natural language streaming comprehension, interactive ink pens
X-Form	Interactive complex form	Performing several kind of data manipulation tasks,	Web forms, Interactive GUI, Special windows
InkML	Graphics	Building data formats used to represent digital ink entered with an electronic pen or stylus; Transferring digital ink data among devices and software components; Storing hand-input traces for handwriting recognition, signature verification, and gesture interpretation	Providing a simple and platform-neutral data format to promote the interchange of digital ink among different kinds of software and hardware applications and services, Information entry and manipulation, Use of applications, Use of interactive services, Semantic interpretation of gestures, Multi-touch interaction
VOICEXML	Voice	Speech recognition, Speech synthesis, Control of voice recognition capabilities through grammars, Speech synthesis control	Recognition of spoken input and DTMF input, recording spoken input, and controlling dialogue flow; Transferring and disconnecting telephone calls; Synthesizing speech and audio files; Identifying exactly which data to submit to the server in IP connectivity
SSML	Voice	Multimedia streaming management	Controlling the pronunciation, tone, inflection, and other characteristics of spoken words; Capturing text speed, volume, inflection, and prosody
SALT	Voice	Integration of speech and video	Integrating speech and telephone interfaces with Web applications and services, Supporting speech input and output in visual pages

deal with each XML-based dialect, the specific supported functions, and applications and services in Web and mobile environments.

FUTURE TRENDS

Control and interaction with the environment providing different services, and the new need of different devices such as personal computers, mobile phones, PDAs, TVs, and so on, will improve development of standard solutions and languages. In particular a greater development of a new generation of mobile devices, with different interaction modes according to their pervasive use and services, will produce an increasing interest about the development of the XML-based languages about different interaction mode and integration modes.

CONCLUSION

The emergence of new devices for human-computer interaction and communication characterized by various sizes of displays, voice-based interaction, new devices for handwriting, and character input are all elements that require the user interface to be portable and to support multimodal interaction. That is, XML portability and adaptation to the different devices provide a solution with the different languages that were developed.

This article provides an overview of the evolution of the multimodal XML-based languages used for mobile applications and services. The main focus is on specific applications of each language that are available on the Web environment and mobile devices.

The use of XML and XML-based languages allows solving multimodal problems in several application fields.

Special attention has to be considered for the activity of the World Wide Web Consortium, which is devoted to developing innovative XML technologies (that are standard de facto) in order to make the Web the focal element for data interchange, interoperability, and accessibility to the information according to the different interaction modes and with the different devices.

REFERENCES

Andersson, O., Axelsson, H. et al.(*2003). Mobile SVG profiles: SVG Tiny and SVG Basic.* W3C Recommendation.

Bals, K. (2005). Using XSL, XForms and UBL together to create complex forms with visual fidelity. *Proceedings of the XML Conference.*

Branco, P. (2001). Challenges for multi-modal interfaces towards anyone anywhere accessibility: A position paper. *Proceedings of the Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly.*

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2004). *Extensible Markup Language (XML) 1.0* (3rd ed.). W3C Recommendation.

Burnett, D., Walker, M., & Hunt, A. (2004). *Speech Synthesis Markup Language (SSML) version 1.0.* W3C Recommendation.

Chamberlin, D., & Goldfarb, C. F. (1987). Graphic applications of the Standard Generalized Markup Language (SGML). *Computers & Graphics, 11*(4), 343-358.

Chatty, S., Lemort, A., Sire, S., & Vinot, J.L. (2005). Combining SVG and models of interaction to build highly interactive user interfaces. *Proceedings of the 4th Annual Conference on Scalable Vector Graphics (SVG Open 2005)*, Enschede, The Netherlands.

Cisco Systems, Comverse, Intel Corporation, Microsoft Corporation, & Philips Electronics. (2002). *SALT.1.0.doc©.* SpeechWorks International Inc.

Hunt, A., & McGlashan, S. (2003). *Speech recognition grammar specification version 1.0.* W3C Proposed Recommendation.

Kvale, K., Warakagoda, N. D., & Knudsen, J. E. (2003). Speech centric multimodal interfaces for mobile communication systems. *Telektronikk.*

Lucas, B. (2000). VoiceXML for Web-based distributed conversational applications. *Communications of the ACM, 43*(9), 53-57.

Marrin, C., Myers, R., & Aktihanoglu, M. (2005). *Emma: An extensible multimedia architecture. Emma White Paper.* Retrieved from emma.3d.org

Mohamed, K. A., Belenkaia, L., & Ottmann, T. (2004). Post-processing InkML for random-access navigation of voluminous handwritten ink documents. Proceedings of WWW, Alternate Track Papers & Posters (pp. 266-267).

Musciano, C., & Kennedy, B. (2003). HTML & XHTML: The definitive guide. *Inf. Res., 8*(2).

Okazaki, N., Aya, S., Saeyor, S., & Ishizuka, M. (2002). A multimodal presentation markup language MPML-VR for a 3D virtual space. *Proceedings of the Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges (in conjunction with HF2002 and OZCHI2002)*, Melbourne, Australia.

Paternò, F. (2004). Multimodality and multi-device interfaces. *Proceedings of the W3C MMI Workshop on Multimodal Interaction.*

Solon, A., Mc Kevitt, P., & Curran, K. (2004). *Mobile multimodal dynamic output morphing tourist systems.* Intelligent Multimedia Research Group, University of Ulster, Magee Campus, Northern Ireland.

Zong, Y., Dohi, H., & Ishizuka, M. (2000). *MPML 2.0e: Multimodal presentation markup language supporting emotion expression*. Department of Information and Communication Engineering, School of Engineering, University of Tokyo, Japan.

KEY TERMS

Extendible Form (XForm): A W3C markup language representing the next generation of forms for the Web and mobile systems, used to describe general user interfaces.

Extensible Multimodal Annotation Markup Language (EMMA): An XML-based W3C recommendation used for information automatically extracted from the user's input. It focuses on annotating the interpretation of input information.

Multimodal Presentation Markup Language (MPML): A W3C markup language for use in systems that provide semantic interpretations for a variety of inputs.

Ink Markup Language (InkML): A W3C language to describe ink data acquired with an electronic pen or similar device.

Scalable Vector Graphics (SVG): A W3C markup language describing a two-dimensional vector graphic. It enables shapes, raster graphics images/digital images, and text.

Speech Application Language Tag (SALT): A W3C markup language used for adding voice recognition to HTML and XHTML files.

Speech Synthesis Markup Language (SSML): An XML-based W3C recommendation for speech synthesis.

Synchronized Multimedia Integration Language (SMIL): A W3C-recommended XML-based markup language for multimedia presentations.

VoiceXML: The W3C standard XML format for specifying voice dialogues between a human and a computer.

Index

Symbols

[mK]M 521
[mKM] 521
2G (see second generation)
3APL-M 244
3G (see third generation)
  commercial deployment 940–946
  mobile technology 632
  network 778
3GPP (see 3rd Generation Partnership Program)
3rd Generation Partnership Program (3GPP) 876
4G (see fourth generation)
7DS (see seven degrees of separation)
802.11 WiFi 90

A

A-GPS (see assisted-GPS)
A/I (see atomicity and isolation)
AAA (see authentication authorization accounting)
AAC (see augmentative-alternative communication)
academia 365
academic PN (AcPN) 1, 7
access 1004
  control 227 , 832
  discovery 812
  layers 624
  network 873
  point (AP) 172, 195, 201
  selection 812
accessibility 9, 27
accounted attributes 334
accoustic data channel 15
ACK (see acknowledgement)
acknowledgement (ACK) 395
ACL (see asynchronous connectionless)
AcPN (see academic PN)
active
  credibility 25
  element (sensor) 868
  server page (ASP) 912
  registry (AR) 438
Active Bat location system 863
ad-hoc 693
  computer network 393, 395, 397
  network 397, 1029
  on demand distance vector (AODV) 427
  way 878
adaptive
  multi-rate voice codec (AMR) 726
  real-time application 24
  signal processing 356, 620
adaptivity 243
ADC (see analogue-to-digital converter)
address notification 257
addressing 882
adjacent relationship query 791
admission controller 196
ADSL 418
advanced
  mobile phone system (AMPS) 266
  video coding (AVC) 672
adventure game 929
advertising 886
  service 72, 75
AeroScout 774

Copyright © 2007, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

agent 218, 386, 434, 831
  -based inter-tree handover 544
  freeze 431–432
  migration pattern 994
  of LSIS 790
  platform (AP) 243
  pre-activation 432
  receptionist 431
  transport 431, 433
aggregator 503
AGM (see anycast group membership)
Agora 436
AHP (see analytic hierarchy process)
AKC (see area key controller)
AKE (see authenticated key exchange)
alert services 713
allowable operations 950
ALM (see application-level multicast)
ALN (see application-level networking)
ALSAR (see application load sensitive anycast routing method)
ambient
  intelligence 243
  Wood project 525
American
  customer satisfaction model (ACSM) 143
  Foundation for the Blind 569
AMR (see adaptive multi-rate voice codec)
analogue-to-digital converter (ADC) 19, 669
analysis 791
analytic hierarchy process (AHP) 12, 28, 183, 376
angle of arrival (AOA) 511, 942
animation 672
annotation and metadata 602
anonymity 840
antenna beamforming 731
anthropology 519
anycast
  group membership (AGM) 54
  open shortest path first (AOSPF) 52
any source multicast (ASM) 541, 592, 594
AOA (see angle of arrival)
AODV (see ad-hoc on demand distance vector)
AoI (see areas of interest)
AOSPF (see anycast open shortest path first)
AP (see access point)
API (see application programming interface)
application 459, 998
  -level
    multicast (ALM) 395, 397
    network (ALN) 397
    networking (ALN) 395
    security 801
  broker 386
  load sensitive anycast routing method (ALSAR) 52
  module 342
  programming interface (API) 250–251, 344, 514, 754
  server 388, 727
  service provider (ASP) 1006
  system 181
  tunneling (see also port forwarding) 537, 540
AR (see augmented reality)
architecture 899
  -specific pattern 745
area key controller (AKC) 834
areas of interest (AoI) 390
ARPU (see average revenue per user)
artificial
  limb 651
  neural network 69
Asia-Pacific 906–911
ASM (see any source multicast)
ASP (see application service provider)
ASR (see automatic speech recognition)
assertion 59
assisted-GPS (A-GPS) 511, 770
assistive technology (AT) 352, 356, 403, 616, 620
associate identification (AID) 1025
associating 632
asymptotic cost 738
asynchronous
  connectionless (ACL) 344
  transfer mode (ATM) 142
AT (see assistive technology)
ATM (see asynchronous transfer mode)
atomicity and isolation (A/I) 949
attributes 60
audio memo 757
augmentative and alternative communication (AAC) 352, 356, 574, 616, 620
augmented reality (AR) 207, 212
authenticated key exchange (AKE) 1021
authentication 581, 708, 840–841, 970
  authorization accounting (AAA) 802
  header 255
auto-configuration 253
automated speech recognition (ASR) 646
automatic
  parking equipment 403
  profile comparison 251
  speech recognition (ASR) 677, 1054
autonomic
  computing system 64
  manager 64
autonomy 243
availability, 622
  analysis 238
AVATON 387

average
  perception 415
  revenue per user (ARPU) 138

B

backward
  handoff 969
  secrecy 232 , 837
Bakhtin 519
bandwidth 124, 268, 459, 489, 627, 805
  probing 788
  reserved per request 983
  utilization 788
banner advertising 636
barcode 898
  mobile coupon 643, 888
baseline profile (BP) 592
base station (BS) 172, 739, 743, 810, 950, 980
basic
  object adapter (BOA) 162
  sign-off 235, 951
batching 369
baton handover 944
Bayesian network (BN) 865–868
BB (see broadband)
belief-desire-intention (BDI) 244, 315
BGCF (see breakout gateway control function)
bi-directional tunneling 543
binary
  -phase-shift keying (BPSK) 273
  runtime environment for wireless (BREW) 309, 503
binding 256, 259
  acknowledgement 259
  cache 257
  request 258
  update (BU) 257–259
    message 779
biometric
  security 561
  user authentication 841
Blackberry 572
blended learning 905
blind 655
block-based method 890
blog 901
blogging 905
Bluetooth, 118, 249–250, 272, 341, 394, 488, 507
    749, 804–809, 820, 825, 839, 1011
  device address 825
  link 344
  Network Encapsulation Protocol (BNEP) 805
BN (see Bayesian network)
BOA (see basic object adapter)
board game 929
body motion 406
  control system 403
bookmarking 79
bootstrap agent transport 433
border node 328
BP (see baseline profile)
BP-Services 84
BPSK (see binary-phase-shift keying)
Braille 570
brain computer interfacing 68–70
branching 989
break 603
breakout gateway control function (BGCF) 874
BREW (see binary runtime environment for wireless)
bricks-and-mortar store 312
bridging 1044
broadband (BB) 458
broadband wireless access (BWA) 921
broadcast 402
  communication 399
  environment 159
broadcasting 400
broker pattern 745
browser 78–83
  -based architecture 646
  phone 476
  software 34
browsing 346, 488
  behaviour 603
BS (see base station)
buffering 970
Build-a-PC 323
built-in wireless technology 178
business
  -to-business (B2B) commerce 33
  -to-consumer (B2C)
    commerce 33
    environment 334
  logic layer 262
BWA (see broadband wireless access)

C

CABAC (see context-adaptive binary arithmetic coding)
CAC (see call admission controller)
cache
  coherence
    (see also invalidation strategy)
  invalidation strategy 104, 107
  management strategy 107, 159, 854
  replacement 750–751
caching 107, 154, 159, 173–177, 369, 854
call
  admission controller (CAC) 197, 201

  -down system 394
  session control function (CSCF) 874
caller/cellular authentication and voice encryption (CAVE) 691
CAMP (see core-assisted mesh protocol)
capacity 357
CAR (see correspondent anycast resopnder)
card emulation 707
care-of-address (CoA) 53, 255
cargo tracking 660, 907–908
Carnival 516–517, 519
carrier
  sense (CS) 329
    medium access with collision avoidance (CSMA-CA) 273, 276
  sensor multiple access (CSMA) 329
cartographic data 390, 392
cascading style sheet (CSS) 377, 1054
CAVE (see caller/cellular authentication and voice encryption)
CAVLC (see context-adaptive variable-length coding)
CBSD (see component-based software development)
CDC (see connected device configuration)
CDMA (see code decision multiple access)
CDR (see common data representation)
cell 194, 739, 743
  -ID 770
  of origin (COO) 511
cellular
  architecture 1038
  phone 895
    cloning 688, 691
    tower 691
  technology 415
conceptualization 284
centralized
  P2P 89
  service provision 880
central processing unit (CPU) 318, 585
central scheduling
  configuration message (CSCF) 927
  message (CSCH) 927
certification authority (CA) 582
channel
  allocation 732
  state information (CSI) 166
character
  agent system 1052
  control 1052
chat 778
check-out mode 234, 951
childcare 660
cHTML (see compact HTML)
civic structure 935–936
classification 377
CLDC (see connected limited device configuration)
CLIC 522
client 219, 388
  -server
    application 260
    network 954
  site 790
  /server model 397
  -side handheld
    computing 302
    programming 309
clone 59
cluster 7
  gateway 738
clustered model 948
clustering (grouping) 203, 205, 347
CN (see correspondent node)
CoA (see care-of-address)
code
  decision multiple access 329
  division multiple access (CDMA) 96, 456, 503, 946
  mobility 576, 578, 580
Code Division Multiple Access 2000 (CDMA 2000) 535
cognitive
  dimensions of notations 14
  load 1002
collaborative
  learning 901
  signal processing 204–205
collection layer 864–865
collision
  avoidance 329
    scheme 924
combinatorial optimization 683
common spatial information services interfaces (CSISI) 790
comma separated value (CSV) 366
commerce 38
common data representation (CDR) 160
communicating
  sequential process (CSP) 580
    for Java (JCSP) 576
communication 901
  availability 234
  cost 230, 238
  distance 1046
  efficiency 835
  network 369
community
  of practice (CoP) 529, 532
  support 714
Compact Flash (CF) card 770
compact HTML (cHTML) 376

compatibility 413–414, 418
  test 721
compensating mode 948
complements 12
complexity 413
  analysis 738
component
  -based software development (CBSD) 58, 61
  configurator 747
  model 61
composite attack 829
computation
  cost 231
  efficiency 836
  tree logic (CTL) 994
computational resources 877
computed
  radiographic (CR) 533
  tomographic (CT) 533
computer
  -based cognitive tool 318, 327
  -supported collaborative work (CSCW) 634, 934
computing system 734
concept match 377
concurrency control 234
conference
  communication 589
  signaling 590
conferencing service provider 152
confidence score 337
confident 573
confidentiality 581, 840
congestion control 812
connected
  device configuration (CDC) 305, 980
  limited device configuration (CLDC) 305, 793, 980
constrained optimization 683
construction 519
  rebroadcast 751
consumer 9, 312
  -centric approach 461
  demand 78
consuming services 72, 75
content
  -based taxonomy 928
  -sharing software 960
  adaptation 75
  management 118
  optimization 788
  page 366
  piracy 534
  provider (CP) 392, 982, 986
  server 388
  transformation 123
context
  management 800
context 7, 118, 398, 402, 716, 722
  -adaptation 124
  -adaptive
    binary arithmetic coding (CABAC) 592
    variable-length coding (CAVLC) 592
  -aware
    application 137
    computing 510, 769
    network system 138
    security 800
  -enabled adaptive service 140
  -sensitive
    profile 251
    service logic 141
  awareness 124, 130, 711
  information 116
  metadata 116–118
contiguity 517
continuity 57
continuous query 665
  table (CQT) 662
contrast sensitivity functions (CSF) 760
control 882
  list 1030
  module 354 , 618, 630
  plane 873
controlling PoC function 727
convenience 284, 290, 398
convergence technology 149–153
COO (see cell of origin)
cookies 79
cooperative
  caching 154, 159
  data dissemination 750
CoP (see community of practice)
CORBA® 436
core-assisted mesh protocol (CAMP) 395
corporate social responsibility (CSR) 469
correspondent
  anycast responder (CAR) 54
  node (CN) 256–257
cost concerns 181
coverage analysis 791
CP (see content provider)
CPU (see central processing unit)
CQT (see continuous query table)
CR (see computed radiographic)
credibility 25
credit card use 416
Cricket 774
  Location Support System 863
crops management 763

cross-layer protocol engineering 166
crossing 739
crossword puzzle 323
cryptography 831, 979
CS (see carrier sense)
CSCF (see call session control function)
CSCH (see central scheduling)
CSCW (see computer-supported collaborative work)
CSISI (see common spatial information services)
CSMA (see carrier sensor multiple access)
CSMA-CA (see carrier-sense medium access with collision avoidance)
CSP (see collaborative signal processing; client/server paradigm)
CSP (see communicating sequential processes)
CSS (see cascading style sheet)
CSV (see comma separated value)
CT (see computed tomographic)
culture 313
customer
  priorities 500
  profiling 99
  service 442, 473
  support 500
customization 463, 652
cyclic redundancy code (CRC) 840

D

DAA (see data access agent)
DAB (see digital audio broadcast)
DAD (see duplicate address detection)
DARPA (see Defense Advanced Projects Agency)
data
  -transfer rate 534
  access 906
    agent (DAA) 947
  and information readiness 484
  attack 828
  caching 172–177
  collecting infrastructure 203–205
  collection 563
  compression 535
  dupression 121
  fusion 209, 676, 868
  integrity 840
    protection protocol 827
  layer 262
  management 536
  sharing 181, 219
  sources description information (DSDI) 214
  usefulness 180–181
database
  management system (DBMS) 303, 974
  query system 369
DBMS (see database management system)
DBN (see dynamic Bayesian network)
DB partition 951
DCA (see dynamic channel allocation)
DCF (see distributed coordination function)
decision support 12, 28
decomposition 791
Defense Advanced Projects Agency (DARPA) 427
degrees-of-freedom (DOF) 212, 656
delay jitter 24
delivery context 14, 26, 30, 380
demodulation 17
denial of service (DoS) 534, 839, 1024, 1028
Department of Defense 32
dependability 233, 800
description logics (DL) 44, 50, 377
Design for All 861
desktop PC 3
destination options header 255, 258
detected attribute 334
detection 204
developer 503
device heterogeneity 75
dexterous hand 653–654
DFSK (see differential FSK)
DHCP (see dynamic host configuration protocol)
DHT (see distributed hash table)
diagonal replication grid (DRG) 234
dial-up 1023
diameter 802
DICOM (see digital imaging and communications in medicine)
different handoff scheme 624
differential
  FSK (DFSK) 16
  phase-shift keying (DPSK) 272
diffusion 413
digital
  assistant 246
  audio broadcast (DAB) 455
  battlefield 660
  cellular phone 475
  certificate 434, 827, 831
  extended broadcast 613
  imaging 318
    and communication in medicine (DICOM) 533, 540
  ink 679
  literacy 568
  radiographic (DR) 533
  technique 942
  technology-mediated communication 562
  technology (2G) 999
  video 671
    encoding 889

DII (see dynamic invocation interface)
direct
  migration technique 120
  sequence spread spectrum (DSSS) 273, 276
dirt 519
disability-centered organization 569
discovery 882
  process 209
discrete cosine transform (DCT) 594, 629
display 536
  modification 79
  module 354, 618
distance learning (d-learning) 423
distraction 517
distributed
  client architecture 646
  computing 202, 205
    system (DCS) 436
  coordination function (DCF) 195
  databank 624
  detection and estimation 204–205
  hash table (DHT) 494, 954, 960
  provision 881
  query and search 203–205
  scheduling message (DSCH) 927
  speech recognition (DSR) 645
  target tracking 204–205
  tracking environment 209
DL (see description logics)
DMS (see database management system)
dots per inch (dpi) 670
DOF (see degrees-of-freedom)
dot.com bust 35
download 488–489, 707
DPSK (see differential phase-shift keying)
DR (see digital radiographic)
DRG (see diagonal replication grid)
DSCH (see distributed scheduling)
DSDI (see data sources description information)
DSI (see dynamic skeleton interface)
DSR (see dynamic source routing)
DSSS (see sequence spread spectrum)
DTMF (see dual tone multiple frequency)
dual-slot mobile phone technology 475
dual tone multiple frequency (DTMF) 1053
duplicate address detection (DAD) 259
dynamic
  applications suitability 57
  Bayesian network (DBN) 863, 869
  channel allocation (DCA) 944
  content adaptation 624
  environment 877
  home agent address discovery 53
  host configuration protocol (DHCP) 882
  invocation interface (DII) 161
  routing protocol 425
  skeleton interface (DSI) 162
  source routing (DSR) 427
  trust relations 798
  weighing 651
dynamically discoverable 878

E

e
  -commerce (see electronic commerce)
  -coupon 299
  -government 581
  -learning advancement 419
E-OTD (see enhanced observed time difference)
ease of
  service provisioning 140
  use 414, 418
eavesdropping 1024
EDGE (see enhanced data rates for global evolution)
education 633
EFR (see enhanced full rate)
EG (see event grammars)
EGPRS (see enhanced GPRS)
electric lock interface 1013
electronic
  commerce (e-commerce) 32, 34, 36, 38, 41, 96, 108, 283, 311, 339, 435, 831, 974
  learning (e-learning) 423, 525
  product code (EPC) 183, 819
  serial number (ESN) 691
  service guide 613
  wallet 636–637
embedded system 260
EMMA (see Extensible MultiModal Annotation Markup Language)
employee readiness 485
enabled task 1047
encrypted tunnel 1032
encryption 435
end-to-end QoS 802
end user 7
energy model 333
enhanced
  data for global evolution (EDGE) 535
  full rate (EFR) 726
  GPRS (EGPRS) 726
  observed time difference (E-OTD) 511
enhancement 483
enterprise
  readiness for mobile ICT 486
  transformation 486
entertainment 633
entity 118

entry permit 431
environment
  -specific inter-ORB protocols (ESIOPs) 160
  properties 10
EPC (see electronic product code)
equalization 165
escape from formalism 11
ESIOP (see environment-specific inter-ORB protocol)
ESN (see electronic serial number)
event 604
  control action pattern 747
  grammars (EG) 208
eventing 882
evictor pattern 746
execution sequence 988
experienced credibility 25
experimental verification 787
expertise 13, 25
explicit 378
exploratory factor analysis (EFA) 299
exposed-terminal problem 168
extended profile (XP) 592
extendible form (XForm) 1056
eXtensible
  HyperText Markup Language (XHTML) 376, 1050
  Markup Language (XML) 89, 142, 376, 456, 675, 912, 1050
  MultiModal Annotation Markup Language (EMMA) 1052, 1056
extensible
  authentication protocol (EAP) 1031
extension header 254
eye tracker (ET) 677

F

facade pattern 746
FAR (see furthest away replacement)
fast
  fourier transform (FFT) 16, 19
  learning 526
fault tolerance (FT) 537, 622, 925
Federal Communications Commission (FCC) 769
FCFS (see first come first serve)
FDD (see frequency division duplex)
FDDI (see fiber distributed data interface)
FDMA (see frequency division multiple access)
Federal Communications Commission (FCC) 569
federated identity management 876
FFD (see full function device)
FFT (see fast fourier transform)
FHSS (see frequency-hopping spread spectrum)
fiber distributed data interface (FDDI) 142
FIFO (see first in first out)
fighting game 928
financial security 285, 290
first
  come first served (FCFS) 447
  generation (1G) game 187
  in first out (FIFO) 447
  -person shooting (FPS) 929
fission 645
fixed
infrastructure 738
  price charging 549, 552
FL (see foreign link)
fleet management 660
flexibility 149, 435, 831
FlexiSPY 557
flooding 395
foreign
  device 7
  link (FL) 256
formal learning 532
forward
forward
  handoff 969
  secrecy 232, 837
forwarder-receiver pattern 745
forwarding zone 395
fourth generation (4G) 733
  game 187
  mobile system 682
fragment header 255
frame rate 24
framework model 370
FreeTV 611
frequency
  -hopping spread spectrum (FHSS) 272, 276
  division
    duplex (FDD) 940
    multiple access (FDMA) 329
  resolution 19
  response 19
  shift keying (FSK) 16–19
Fresnel Zone 1028
FSK (see frequency shift keying)
FT (see fault tolerance)
full function device (FFD) 273
functional module 303
furthest away replacement (FAR) 740
fusion 645
  layer 864–865
future prosecution 978
fuzzy
  expert system 763
  logic 763

G

GALILEO 770
GAMA (see generic adaptive mobile agent)
gambling 488, 554
gamer 928
gaming 294, 914
gaming
GAP (see generic access profile)
gate of the logical network 53
gateway GPRS support node (GGSN) 874
Gaussian frequency shift keying (GFSK) 272
gaze tracking (GT) 677
GDBS (see global database system)
GDM (see generative domain model)
gender 296–301
general
  design pattern 745
  inter-ORB protocol (GIOP) 160, 164
  message format 430
  packet radio service (GPRS) 142, 194, 503, 505, 510, 515, 726
generative domain model (GDM) 208
generic
  access profile (GAP) 344
  adaptive mobile agent (GAMA) 717
  log adapter (GLA) 64
genetic algorithm (GA) 339, 349
geo-referenced information (GRI) 213
geocast 393, 395, 397
geocasting-limited flooding 395–396
geographical
  content database 388
  database 385
geographic information system (GIS) 129, 137, 219, 856, 907
Geography Markup Language (GML) 134, 137
geolocation information 773
geometric models 857
gesture recognition (GR) 677
GFSK (see Gaussian frequency shift keying)
GGSN (see gateway GPRS support node)
GIOP (see general inter-ORB protocol)
GIS (see geographic information system)
GKS (see group key server)
global
  cache
    hit 155
    miss 155
  computing environment 57
  database system (GDBS) 948
  mobile system (GSM) 504–509
  optimization 683
  positioning
    satellite 327, 739
    system (GPS) 125, 130, 137, 203, 394–395, 462, 511, 637, 662, 769, 856, 885, 898, 907, 1024
  spatial information services (GSIS) 790
  system
    for mobile communication (GSM) 194, 356, 456, 47 6, 515, 620, 726
  transaction (GT) 948
    coordinator (GTC) 949
    manager (GTM) 949
GOEXP (see generic object exchange profile)
GOS (see grade of service)
GOVOREC 573
GPRS (see general packet radio service)
GPS (see global positioning system)
grade of service (GOS) 446
graphic interchange format (GIF) 670
graphics 671
GRI (see geo-referenced information)
Grid
  index information service (GIIS) 437
  information resource service (GRIS) 437
grid service 77
group
  communication 227
  key
    management algorithm 228, 833–834, 834
    server (GKS) 228
GSM (see global system for mobile communication)
GSIS (see global spatial information services) 790
GT (see global transaction)
GTC (see global transaction coordinator)
GTM (see global transaction manager)
guest 998

H

HA (see home address)
HA (see home agent)
hacker 561, 1023, 1029
hand
  controller 654
handheld 611
  computing 309
  device 908
  firewall 840
  terminal (HHT) 894
handoff 230, 621, 834, 837
  user
    list 230
handover 967
hands-free 675
  accessory 571
handwriting recognition (HR) 646, 677, 755

harmonic distortion 19
hash function 1021
  attack 688
hashing 961, 965
HCI (see human-computer interaction)
head controller 656
headhunter (HH) 438
health promotion 634
healthcare 504
  challenge 1010
  organization 1004
hearing accessories 571
helper application 625
heterogeneous
  device 387
  multimedia network 796
  wireless technologies 796
HHT (see handheld terminal)
hidden-terminal problem 168
hierarchical structure (see also tree structure) 832
high
  -end
    consumer device 978
    device 305
  -featured mobile phone device 4
  available server cluster 537
histogram 890
HL (see home link)
HLS (see homeland security)
HN (see home network)
home
  address (HA) 53, 256–257
  agent (HA) 53, 256
  link (HL) 256
  location agent 969
  network (HN) 53
  PoC server 727
  subscriber server (HSS) 874
homeland security (HLS) 393
hop-by-hop options header 254
horizontal
  handover 967
  interaction 165
  mobile business 442
host 998
  auto-configuration 53
hotspot 44
HSS (see home subscriber server)
HTML 671
HTTP (see hypertext transfer protocol)
human
  -centered LBS 856
  -computer
    interaction (HCI) 68, 676, 935
    interface 313, 403
  -machine interaction 651–659
  behavior 652
  interfacing 804
  properties 9
  visual system (HVS) 760
hybrid
  coupling 810
  data dissemination 749
  intelligent system (HIS) 763
  model 857
  mobile device 138
  P2P 89
Hyper-Text Markup Language (HTML) 456
HyperLan 805
hypertext transfer protocol (HTTP) 376, 1023

I

i-menu 296
i-mode 643, 871, 888,
ICT (see information and communication technology)
ID (see identification number)
  card 477
identification 430
  number (ID) 826
identity
  management 798
  vector 967
IDL (see Interface Definition Language)
IDT (see innovation diffusion theory)
IGO (interactive graphing object) 322
iGrocer 312
image 414, 418
  layer 389
  resolution 535
  transmission 535
imaging 670
IMC (see integrated marketign communications)
IMEI (see international mobile equipment identifier) 692
immutability 522
implementation 829
  neutrality 878
improved mobile telephone system (IMTS) 266
IMS (see interactive multimedia system)
IMS (see IP multimedia subsystem)
incentive-based marketing 934
inconsistency ratio (IR) 180
incremental evaluation 665
individual view 935
indoor
  navigation ontology (INO) 858
  positioning 859, 861
  spatial model 859
informal learning 532

information
  and communication technology (ICT) 33, 319, 466, 481, 516, 519, 563, 568
  availability requirement 191
  security requirement 191
  services 789, 795
  technology (IT) 283, 466
infotainment 593
infrastructural pattern 745
infrastructure
  element 869
  sensor 865
initial link 87
initialization vector (IV) 1022
initializing vector (IV) 1030
Ink Markup Language (InkML) 1053, 1056
InkML (see Ink Markup Language)
innovation-diffusion theory 33
innovation diffusion theory (IDT) 33, 894
innovativeness 413
INO (see indoor navigation ontology) 858
input module 353
inspection 12, 28
input module 617
instantaneity 517
instant messaging (IM) 488, 778, 967
Institute of Electrical and Electronics Engineers (IEEE) 252
intangible 327
integrated
  marketing communications (IMC) 885
  service digital network (ISDN) 142, 504, 533
integration manager 645
integrity 581
intelligent
  agent 561, 1000
  decision support system (IDSS) 386
  LBS 856
  sensory system 653
inter
  -provider relationship 798
  -symbol interference (ISI) 940
interaction
  manager 644
  services 789, 795
  trajectory 935–936
interactive 398
  catalog 345
  graphing object (IGO) 322
  multimedia system (IMS) 341–344
  presentation 1052
  TV (iTV) 633
  voice response (IVR) 1054
interactivity 928
interface
  category 498
  repository (IR) 162
Interface Definition Language (IDL) 160, 164
international
  law 393
  mobile equipment identifier (IMEI) 692
International Telecommunication Union—Telecommunication Standardization Sector (ITU-T) 594
Internet 78, 298, 489, 507, 553, 612
  -based learning 419
  -enabled mobile handset 639
  inter-ORB protocol (IIOP) 164
  packet exchange (IPX) 1023
  protocol (IP) 51, 504, 537, 700
    -based
      network 20
      wireless communication system 730
    -multicast 394
    multimedia subsystem (IMS) 138, 724, 801–802
  service provider (ISP) 35
  technology 416
  usage 78–79
interoperability 149, 463
interoperable object reference (IOR) 160
interpersonal communication 291
intuitively-correct 857
invalidation
  report (IR) 107
  strategy (see also cache coherence) 849
inventory 908
inverse replication (IR) 964
involvement 284, 290
IOR (see interoperable object reference)
IP (see Internet protocol)
  address 253, 980
  mobility 253
iPod 365, 489
IPv4 253
IPv6 253
IR (see inconsistency ratio)
IR (see interface repository)
IR (see invalidation report)
IR (see inverse replication)
ISDN (see integrated services digital network)
ISI (see inter-symbol interference)
itinerary attack 828
ITU-T (see International Telecommunication Union—Telecommunication Standardization Sector) 594
iTunes 78
iTV (see interactive TV)
IVR (see interactive voice response)

J

J2ME (see Java 2 Platform, Micro Edition)
Java
  media framework (JMF) 915
  virtual machine (JVM) 914
Java 2 Platform, Micro Edition (J2ME) 193–194, 269, 309, 367–368, 503, 667, 668, 795
  security model in its standard edition (J2SE) 979
JCSP (see communicating sequential processes for Java)
JD (see joint detection)
jini 77, 882
Joey transaction (JT) 948
joint detection (JD) 940
Joint Video Team (JVT) 594
JPEG 670
JPEG2000 670
JT (see Joey transaction) 948
just-in-time learning 526, 905
JVT (see Joint Video Team)

K

k
  -nearest neighbor (k-NN) 660–665
  -NN (see k-nearest neighbor) 660–661
  -shortest paths
    problem 861
    searching algorithm 859
Kalman filter 212
kangaroo transaction (KT) 947
KDC (see key distributor center)
Keitai 296
KEK (see key encyption key)
kernel 342
key
  attack 828
  distributor center (KDC) 832
  encryption key (KEK) 227, 232, 832–833, 838
  healthcare system input 1010
  management 1020
    algorithm 232, 838
  mapping 961
  seed negotiation protocol 826
keyword-based language 370
killer application 856
Kismet 1024
KMS (see knowledge management system)
knowledge
  management system (KMS) 376
  readiness 485
  representation 380
  worker 773
Knowledge Query and Manipulation Language (KQML) 429
KQML (see Knowledge Query and Manipulation Language)
KSACI 245
KT (see kangaroo transaction)

L

L2CAP (see logical link control and adaptation protocol)
lab information system (LIS) 1005
LAN (see local area network)
laptop 3, 476, 576, 754, 757, 1011
large mobile host (LMH) 950
lateration 394
lazy acquisition pattern 746
LBM (see location-based multicast)
LBS (see location-based service)
LCR (see low chip rate)
LDD (see location-dependent data)
LDIS (see location-dependent information services)
LDQ (see location-dependent query)
leadership readiness 485
LEAP 244
learning 318, 900
  management system (LMS) 318, 423
  support 900
leasing pattern 746
least
  frequently used (LFU) 746
  recently used (LRU) 135, 740, 746
  risk path 857
leaving 835
legal context 26
legality 27
liberty alliance 803
LIGLO (see location-independent global names lookup)
limited
  connectivity 713
  ergonomics 713
  resources 399, 713
  usability 399
link
  manager protocol (LMP) 272
  server 476
Linux 366–368, 895
liquid crystal display (LCD) 533, 536
LIS (see lab information systems)
literacy 568
live streaming 778
LKH (see logical key hierarchy)
LMCDS (see location-based multimedia content delivery system)
LMDS (see local multipoint distribution system)
LMH (see large mobile host)
LMP (see link manager protocol)

LMS (see learning management system)
load balancing 53, 925
local
  area network (LAN) 78, 172, 253, 533, 627
  cache hit 155
  information services 53
  multipoint distribution system (LMDS) 455
  spatial information service (LSIS) 790
  spread
    inverse replication (LSIR) 964
    replication (LSR) 964
  transaction (LT) 948
  wireless interface (LWI) 15
Locales
  Foundation 935
  Framework 934
localization 203, 205, 525
location 118, 863
  -aware
    application 863
    content provision 856
    multicast 395
    system 526
  -based
    game service 660
    multicast (LBM) 395
    multimedia content delivery system (LMCDS) 381, 386
    service (LBS) 129, 137, 392, 393, 396, 402, 461, 510, 515, 660, 665, 773–777, 789, 795, 885, 1042
  -dependent
    cache invalidation 105–107
    data
    data (LDD) 393, 739
    information service (LDIS) 743
    query (LDQ) 739, 743, 854
      processing 394
  -independent global names lookup (LIGLO) 85
  -tracking application 561
  estimation system 864
  inference queries 867
  of service 983
  operating reference model (LORE) 394
  server 388–389
logical
  key
    hierarchy (LKH) 227, 232, 835–838
    structure 228, 833
  link control and adaptation protocol (L2CAP) 272
logistics 895, 898
long-term shared key 1021
longevity 522
lookup
  pattern 746
  server 882
loose coupling 359, 810, 878
LORE (see location operating reference model)
low
  -end consumer device 978
  chip rate (LCR) 940
loyalty 906
LRU (see least recently used)
LSIR (see local spread inverse replication) 964
LSIS (see location spatial information service) 790
LSR (see local spread replication) 964
LT (see local transaction) 948
LWI (see local wireless interface)

M

m
  -advertising (see mobile advertising)
  -commerce 345–351
    device (MCD) 38
  -commerce (see mobile commerce)
  -health 1010
  -learning 318
  -payment (see mobile payment) 714
m[KM] 521
MAC (see medium access control)
MAC (see mobile agent communication)
Macromedia Flash 672
MADGIS (see mobile agent-based distributed geographic information system)
MAE (see mobile agent environment)
magnetic resonance (MR) 533
MAI (see motion adaptive indexing)
MAI (see multiple access interference)
main profile (MP) 592
malevolent hackin 534
malicious software 581
MAM (see mobile agent manager)
man
  -in-the-middle attack 1025, 1028, 1031
  -machine interface 403–412
MAN (see mobile agent naming)
management message modified 196
manager 644–645
MANET (see mobile ad-hoc network)
maneuverability 41
manufacturing logistics 896
mapping 723
  candidate 721
MAR (see mobile AR)
MARI (see multi-attribute resource intermediary)
marketing 99–100
  literature 934
MAS (see mobile agent security)

massively multi-player online game (MMOG) 929
MAT (see mobile agent transportation)
maturity 41
maximum transmission unit (MTU) 805
MBR (see minimum bounding rectangle)
MC (see mobile client)
MCOP (see multi-constrained optimal-path problem)
MCU (see multipoint control unit)
mean square error (MSE) 758
MEC (see mobile electronic commerce)
media
  access controller (MAC) 201
  gateway (MGW) 874
  gateway control function (MGCF) 874
  object 390
  player 757
  resource function
    controller (MRFC) 874
    processor (MRFP) 874
  processing 386
  support module 981
  synchronization 1052
mediator 312
medical science 1004
medical sensor 504
medium access control (MAC) 150, 272, 276, 329, 921
member
  join 228, 834
  leaving 834
memory
  -limited mobile device 260–264
  card 839
mental context 116
mesh
  base station (mesh BS) 921
  BS (see mesh base station) 921
  network 981
message
  authentication code (MAC) 840, 1021
  latency 192
Message F (Free) 635, 638
meta-repository (MR) 438
metrics 12, 28
metropolitan area network (MAN) 839
MGCF (see media gateway control function)
MGW (see media gateway) 874
MH (see mobile host)
micro
  -contractor 519
  -payment 33, 460
microcontroller (MCU) 1011
microelectronics 877
Microsoft mobile Internet toolkit (MMIT) 912
microwave multipoint distribution system (MMDS) 455
MIDI (see musical instrument digital interface)
middleware pattern 745
MIDlets 979
MIDP (see mobile information device profile)
migration 119, 988, 998
  design pattern 987
migration
MIMO (see multiple-input multiple-output)
MIN (see mobile identification number)
MIND (see mobile IP-based network development)
minimum bounding rectangle (MBR) 740
MIS (see mobile information system)
MLBG (see mobile location-based game)
mLMS (see mobile learning management system)
MMA (see mobile mote agent)
MMD (see multimedia domain)
MMDBS (see mobile multi-database system)
MMDS (see microwave multipoint distribution system)
MN (see mobile node)
MNC (see multinational corporation)
MNO (see mobile network operator)
MobiAgent 245
MOBIDIS1 1046
Mobile
  Entertainment Forum (MEF) 487
  Magnifier 570
mobile
  access 25
  accessibility 14
  ad-hoc network (MANET) 395, 399, 402, 424, 427, 492, 734, 749–753, 803, 925, 950
  advertising (m-advertising) 398, 402
    (see also wireless advertising, mobile marketing)
  agent (MA) 218–219, 439, 580, 717, 723, 994, 1001
    -based distributed geographic information system (MADGIS) 213, 219
    migration design pattern 987
    communication (MAC) 216
    environment (MAE) 214, 219
    manager (MAM) 216
    naming (MAN) 216
    security (MAS) 216
    transportation (MAT) 216
  and wireless terminals (MWTs) 711
  application 9, 375, 744
  AR (MAR) 207
  authorization 15
  banner advertising 638
  billing 547, 552
  broadband 870
    wireless access (Mobile-Fi) 733
  business (m-business) 442
  calendar 442
  channel 577, 580

chat 251
client (MC) 190, 194, 745–746
commerce (m-commerce) 15, 32–37, 38–42, 96, 283, 290, 311, 429, 435, 455, 461–465, 472, 480, 547, 552, 933
communication 393, 589, 700
computing 61, 78, 93, 169, 266, 291, 854, 986
  and commerce 466–471
  environment 849
conferencing support 591
content provider 552
controls 912–920
cooperative caching 750
convergence 138
credibility 26
data
  communication 399
  terminal (MDT) 907
device 71, 77, 1043
e-mail 442
electronic commerce (MEC) 414, 974
entertainment 487–491, 669–674

environment 107, 159, 850, 1021
game 185–189, 928–932
  development 497
  industry 930
gaming 497
GIS 134
health (m-health) 504
healthcare
  business model 1010
  delivery model 1010
  delivery system (MHDS) 505
host (MH) 102, 749
hunter 510, 515
ICT 486
identification number (MIN) 691–692
information
  device profile (MIDP) 305, 420, 793, 980
  system (MIS) 190, 194
Internet 213, 296–301, 460
  access 785
IP 1042
  -based network development (MIND) 733
knowledge management (mKM) 520–524
learning (m-learning) 318, 327, 423, 525–527, 528, 899
  management system (mLMS) 420
location-based game (MLBG) 510, 515
lotteries 553
mail advertising 636
marketing 96–101
  (see also mobile advertising, wireless advertising)
marketplace (m-marketplace) 50
module 1012
mote agent (MMA) 328, 330, 333
multi-database system (MMDBS) 949
multicast 541–546, 592
network
  node (MN) 800
  operator (MNO) 547, 714
  technology 497
  virtual operator (MVNO) 876
node (MN) 53, 255–257
object table (MOT) 663
payment (m-payment) 548, 552, 714, 716
  service 706–710
phone 557, 569–575, 644, 999–1000, 1011
  gambling 553–556
  human-interface system 617
  texting 568
positioning 515
process 580
query processing 855
resource 14
robotic system 403–412
service 716
  provider 143–148
station (MS) 150
support
  station (MSS) 190, 194, 749
  system 102
system 369
telephone 318
television (mobileTV) 611–615
terminal 399, 456, 789, 795
ticketing (m-ticketing) 714
transaction 948
  manager (MTM) 948
user 393
virtual
  community (MVC) 632, 634
  network operator (MVNO) 548
Web 9
  engineering 14, 30
MOBIlearn 528, 532
Mobile Speak 570
mobility 73, 243, 318, 461, 525, 576, 632, 644–650, 769, 906
  management 426, 590, 731
  services 789, 795
mobilization 482
Mobitip 311
modern digital communications technology 562
modification score 336
modular system 196
modulation 276

monitoring 660
  and discovery service (MDS) 437
Morse code 352, 356, 616, 620
MOSPF (see multicast open shortest path first)
MOT (see mobile object table)
motion
  adaptive indexing (MAI) 661
  sensitive bounding boxes (MSB) 661
Moving Picture Experts Group (MPEG) 594
moving object 665
MP (see main profile)
MP3 player 4, 365
MPEG (see Moving Picture Experts Group)
MPML-FLASH (see Multimodal Presentation Markup Language with Character Agent Control in Flash Medium)
MPML-VR (see Multimodal Presentation Markup Language for Virtual Reality)
MPML (see Multimodal Presentation Markup Language)
MRFC (see media resource function conroller)
MRFP (see media resource function processor)
MR (see magnetic resonance)
MSB (see motion sensitive bounding boxes)
MSS (see mobile support station)
MTM (see mobile transaction manager)
MUD (see multi-user joint detection)
multi
  -agent system (MAS) 383, 386
  -attribute resource intermediary (MARI) 335
  -constrained optimal-path problem (MCOP) 981
  -database system 947
  -hop routing 1035
  -layer
    -based taxonomy 929
    mobility 966–973
  -objective optimization 683
  -user joint detection (MUD) 940, 943–946
  -media messaging (MMS) 116
multicast 227, 232, 393–397, 397, 541–546, 838
  listener 542
  open shortest path first (MOSPF) 395
  routing protocol 397
  technology 393
multifunctional handheld device 497
multihop communication 427
multimedia 392, 757, 870, 986
  -capable 398
  content
    database 388
    server 390
  data 24
  domain (MMD) 873, 876
  info presentation 389
  message service (MMS) 252, 669, 756–757, 840
  multimodal
    service 1050
    user interface 644–650
  presentation 389
multimediality 675
Multimodal Presentation Markup Language 1052, 1056
  for Virtual Reality (MPML-VR) 1052
  with Character Agent Control in Flash Medium (MPML-FLASH) 1052
multimodal 604
  access 602
multimodality 675–681
multinational
  companies database 886
  corporation 886
  corporation (MNC) 934
multiple
  -input multiple-output (MIMO) 701
  access interference (MAI) 940, 946

multipoint control unit (MCU) 590
Mummy 520
musical instrument digital interface (MIDI) 15, 19
music download 488
mutuality 935–936
MVC (see mobile virtual community)
MVNO (see mobile network virtual operator)
MWT (see mobile and wireless terminal)
MySpace 960

N

naming pattern 746
NAL (see network abstraction layer)
narratives 928
narrowband (NB) 458
national law 393
navigation 861
  algorithm 857
  context 857
  preferences 858
  service 859
  system 125
NB (see narrowband)
NCFG (see network configuration)
NCT (see network connectivity table)
near
  -far problem 168
  -real-time balance management 874
Near Field Communication (NFC) 706
negative update 665
negotiation of bid 983
neighbor
  discovery and management 203, 206
  link 87

  unreachability detection (NUD) 259
neighboring cell 743
NENT (see network entry)
network 392, 981, 986
  -centric
    design viewpoint 873
    model 874
  -level multicast (NLM) 394
  abstraction layer (NAL) 592–595
  configuration message (NCFG) 927
  connectivity table (NCT) 662
  entry message (NENT) 927
  frame 927
  infrastructure 6, 387
  security requirement 191
  topology 986
  survivability requirement 191
networkability 186
networking 207
neutral data format 75
new learning 319, 327
next generation 472
  network (NGN) 142, 796, 803
NGN (see next generation network)
Ninja secure service discovery service (SSDS) 436
NLM (see network-level multicast)
NLOS (see non-line-of-sight)
node 330, 580
  mobility 172
noise 348
  threshold 348–349
nomadic user 576, 589, 591, 595
non
  -deterministic polynomial time (NP) 733
  -formal learning 532
  -line-of-sight (NLOS) 921
  -linear navigation 599
  -quantifiable attributes 334
  -renewable resource 763
  -vital 951
notification service (NS) 711, 716
NS (see notification service)
NUD (see neighbor unreachability detection)

O

OASIS (see Organization for Advancement of Structured Information Standards)
Object Management Group (OMG) 1002

object 604
  adapter 164
  name service (ONS) 178, 183
  -oriented Petri net 994
  request broker (ORB) 160, 164
oblette 532
OBU (see on board unit)
occurrence graph 995
OFDM (see orthogonal frequency division multiplexing)
off
  -line customer 35
  -the-shelf (OTS) 61
offset-quadrature phase-shift keying (OQPSK) 273
old learning 319
on
  -demand
    board unit (OBU) 894, 898
    environment 159
    learning 525
    online community 632, 634
    ontology 723, 862
    session 727
one-time two-factor authentication 533, 540
online
  betting 553
  gambling 553
  merchant 311
  shopping 283, 290
ONS (see object name service)
ontology 50, 339, 380
  universally unique identifier (OUUID) 44
open
  coupling 358
  system
    interconnect (OSI) 165
  system
    model 1030

Open Mobile Alliance (OMA) 463, 613
operational system (OS) 344
operating system (OS) 765, 895, 898, 998
opportunistic
  communication 165
  exploration 345
  scheduling (OS) 165–171
opt
  -in 643
  -out 643
optical character recognition (OCR) 667–668
optimization 791
optimized lifetime 725
options model 39
OQPSK (see offset-quadrature phase-shift keying)
ORB (see object request broker)
organizational security 1011
Organization for the Advancement of Structured Information Standards (OASIS) 876
orthogonal frequency division multiplexing (OFDM) 682, 701

OS (see operating system)
OS (see operational system)
ontology-based context model 718
ontology repository 860
OTS (see off-the-shelf)
OTTFA (see one-time two-factor authentication)
OUUID (see ontology universally unique identifier)
overlay network 738, 965
OWL (see Web Ontology Language)

P

P-CSCF discovery 779
P-PAN (see private personal area network)
P2P (see peer-to-peer)
packet
  data protocol (PDP) 813
  loss rate 22–24
  radio network 427
  routing 257
  transmission 198
Palm OS 304, 309
  Cobalt 304
  Garnet 304
palm pad computer 476
palmtop 757
PAN (see personal area network)
paradigm shift 860
parameterized query 602
Parlay X 876
partial
  global
    indexing (PGI) 742–743
    serialization graph (PGSG) 949
participant 899
passive
  attack 1024
  credibility 25
  distributed indexing 492
password 841
path
  -searching algorithm 857
  -selection rules 858
  optimization 986
pattern 744–748
  extraction 673
  recognition 394
PayTV 611
PC-based online survey 640
PC (see personal computer)
PC (see pervasive computing) 956
PCI (see perceived characteristics of innovating)
PDA (see personal digital assistant)
PDF (see policy decision function) 874
PDU (see protocol description unit)
peak signal to noise ratio (PSNR) 758
pedagogy 532
peer-to-peer (P2P) 2, 84, 89, 159, 395, 397
  architecture 108
  communication 165
  computing 233, 492, 693, 734
  financial transaction 108
  networking 357
pen
  -based interface 754
  computing 757
penetration rates 398
PEP (see performance enhancing proxy)
perceived
  knowledge 25
  quality of service (PQoS) 758
  switching cost 145
perceptual navigation rules 858
performance 27
  comparison 239
  discussion 230
  enhancing proxy (PEP) 785, 788
permission-based marketing 934
personal
  area network 507, 1011
  area network (PAN) 1, 7, 252
  computer (PC) 79, 141
  context 711
  device 7
  device assistant (PDA) 265
  digital assistant (PDA) 10, 38, 79, 138, 154, 172, 2 50, 260, 317, 368, 375, 443, 461, 505, 511, 5 25, 581, 711, 757, 907, 974, 980, 996, 1011, 1015, 1050
    phone 894
  information management (PIM) 249–252, 757
  network (PN) 1, 8
  trusted device (PTD) 581
  virtual environment (PVLE) 530
personality 243
personalization 31, 380, 398, 463, 525
pervasive
  computing (PC) 57, 61, 71, 276, 320, 621, 863, 869, 877–878, 884, 956
  location-aware computing environments (PLACE) 661
pest
  activity 763
  control 763
  management 763
Petri Net 987
PGI (see partial global indexing) 743
PGSG (see partial global serialization graph) 949
physical

  context 116
  navigation rules 858
  network 738
  space 774
Physical Markup Language (PML) 183, 819
PIA-SM (see protocol independent anycast - sparse mode)
piconet 50, 804, 825
picture download 489
pilot
  aliasing 692
    attack 689
  study 641
PIM (see personal information management)
  phone 894
PIM-SM (see protocol independent multicast-sparse mode)
PIN 841
  -code 477
pixel-based method 890
PKI SIM card 840
PLACE (see pervasive location-aware computing environments)
planar structure 822, 825
planned disconnection mode 951
platform-based taxonomy 929
play 604
PML (see Physical Markup Language)
PMP (see existing point-to-point multipoint)
PN (see personal network)
POA (see portable object adapter)
PoC session identifier 727
Pocket PC 304
PoI (see points of interest)
point
  -to-
    multipoint (PMP) 921
    point IP traffic 576
  of sale (POS) 706, 907
points of interest (PoI) 390
policy decision function (PDF) 874
porosity 517
portability 27, 186
portable
  /mobile device 626
  computing 195
  media center 305
  network graphics (PNG) 670
  object adapter (POA) 162
  sensor 865
port forwarding 540
  (see also application tunneling)
position-aware mobile device 561
positioning 869
  method 386
positive update 665
power
  control 731
  management 1035
pre
  -committed transaction 696
  -generation (Pre-G) 187
  -processing of raw data 865
  -roaming notification 432
  -serialization transaction management model 949
premium fee charging 549, 552
presence 1042

presentation layer 262
presenting 882
presumed credibility 25
preview clip 604
price tolerance 145–146
primary notation 11
printer 4
privacy 75, 117, 399, 462–463, 477, 773–777, 1002
  (see also security)
  concerns 713
private
  personal area network (P-PAN) 7
  space 557
pro-motion model 949
proactivity 243
process readiness 484
processing kernel 1013
procurement activity 896
producer 9
product
  brokering 340
  catalog 345
  manager 393
productivity 443, 473, 839, 906
profile information 251
profiling 99
programming languages 915
propitient multi-agent system 957
protocol 435, 831
  data unit (PDU) 584
  description unit (PDU) 352
  independent
    anycast - sparse mode (PIA-SM) 52
    multicast-sparse mode (PIM-SM) 395
  of use 59
proximity 117, 394
proxy 77, 235, 626
pseudo-transaction 236
PSTN (see public switched telephone network)
public 702
  -switched telephone network (PSTN) 139, 142

  - key
    - infrastructure (PKI) 581–588
    - interface SIM (PKI SIM) 840
  - regulation 886
  - space 557
- publisher 503
  - -subscriber pattern 745
- pull 400, 402, 879
  - -based
    - environment 107, 749
    - service provision 880
  - -type advertising 635
  - model 98
- pure P2P 89
- push 400, 402, 879
  - -based
    - environment 107, 749
    - service provision 879
  - -to-talk over cellular (PoC) 724
  - -type advertising 635
  - message 716
  - messaging service 643, 888
  - model 98
- PVLE (see personal virtual environment)

Q

- QFD (see quality function deployment)
- QI (see query interface)
- QoS (see quality of service)
- quadriplegic 403
- quality 14, 31, 758–762, 1004
  - attributes 11, 27
  - function deployment (QFD) 12, 28, 376
  - of service (QoS) 1, 140, 149–150, 195, 201, 207, 379, 425, 621, 624, 627, 701, 729, 746, 778, 803, 804–809, 906, 923, 982, 1042
    - routing 986
  - score 860
- quantifiable attribute 334
- queries generated summary 604
- query
  - decomposer and coordinator 790
  - interface (QI) 214
  - language 369
  - lifetime 738
  - processing 851
- queue manager 197
- queuing 739

R

- R&D (see research and development)
- RA (see router advertisement)
- RADAR 774
- radio
  - access 701
    - network (RAN) 361
    - technology (RAT) 149, 810
  - frequency (RF) 942
    - identification (RFID) 178, 183, 507, 819, 856, 864
  - network design 946
  - resource management (RRM) 425, 729
- radiology information systems (RIS) 1005
- RAID (see redundant array of inexpensive disks)
- RAM (see random access memory)
- random
  - access memory (RAM) 163
  - network 983
- range
  - -incline finder 653
  - -inclination tracer 406
  - query 665
- rapid
  - development 102
  - prototyping 291
- raster data processing engine 389
- RDF (see resource descriptiopn framework)
- reachability 398
- reactivity 243
- read one write all (ROWA) 234
- real-time
  - protocol (RTP) 726
  - requirement 589
  - strategy (RTS) 186
  - tracing 791
  - traffic 626
    - management 622
  - transfer protocol (RTP) 263
  - transport protocol (RTP) 595
- Real Audio 265
- reasoning 862
- received signal strength (RSS) 863
- recognition algorithm 17
- recognizer parameters 18
- recommender system 345
- recommending agent 109
- recycling logistics 896
- redefinition 483
- redirection 1025
- reduced function device (RFD) 273
- redundant
  - array of inexpensive (or identical) disks (RAID) 540
  - array of inexpensive disks (RAID) 533
  - recoding 11
- reflection 60, 164
- rehabilitation robotics 653
- rekey 834
- relative advantage 413–414, 418
- relaxed check-out mode 951

reliability 27
remote
  authentication dial-in user service (RADIUS) 803
  control object 882
  subscription 543
replication 965
reputed credibility 25
research and development (R&D) 700
reshapement 483
resource
  description framework (RDF) 377
  management scheme 812
  readiness 485
response rate 640
results demonstrability 414, 418
retransmission timeout (RTO) 785, 788
returning mobile agent 579
RF (see radio frequency)
RFD (see reduced function device)
RFID (see radio frequency identification)
ring structure 822, 825
RIS (see radiology information systems)
risk 109, 857
  set 111
riskiness value 110
roaming permit 431
robust security network 1032
robustness 27
rogue wireless gateway 1028
role playing game 186, 928
round-trip time (RTT) 785, 788
route
  analysis 791
  optimization 253–259
router
  advertisement (RA) 256–257
  discovery 255, 257
  solicitation 255, 257
routing 394, 807, 965
  algorithm 859
  header 254
  identification 807
  mechanism 732
  metrics 925
  overhead 925
ROWA (see read one write all)
RRM (see radio resource management)
RSS (see received signal strength)
RTO (see retransmission timeout)
RTP (see real-time protocol)
RTP (see real-time transfer protocol)
RTP (see real-time transport protocol)
RTT (see round-trip time)
running task 1047

S

S-MAC (see sensor MAC)
SA (see signal attenuation)
SAFE (see secure roaming agent for e-commerce)
sales logistics 896
SALT (see speech application language tags) 1053
SAML (see Security Assertion Markup Language)
sampling frequency 19
satellite 508
  -based augmentation systems (SBAS) 771
  communication 907
scalability 925
scalable
  vector graphics (SVG) 376, 671, 1050, 1056
  video coding (SVC) 591–592, 595
scatternet 276, 825
Scavenger Hunt (SH) 292
scene analysis 394
schedule frame 927
scheduler 197
SCO (see synchronous connection oriented)
SCP (see servers/client paradigm)
SDAP (see service discovery application profile)
SDDB (see service discovery database)
SDK (see software development kit)
SDK (see software development toolkits)
SDMA (space division multiple access) 944
SDP (see service delivery platform) 870
SDP (see service discovery protocol)
SDP (see session description protocol)
SDP (see sophisticated service discovery protocol)
secondary notation 11
second generation (2G) 456
  mobile telephony 503
  wireless system 96
secure
  authorization 800
  multimedia service access 799
  roaming agent for e-commerce (SAFE) 429
  service discovery 799
  shell (SSH) 842
  socket layer (SSL) 383
security 6, 75, 109, 192, 313, 426, 435, 462, 477, 505, 624, 633, 831, 839–848, 968, 1002, 1011, 1022–1027
  (see also privacy)
  analysis 828
  policy 1028
Security Assertion Markup Language (SAML) 378, 872, 876
segment 604
segmentation and reassembly (SAR) 805
self-independence 569

semantic
  caching 850, 855
  distance 43
  interpretation (SI) 1054
  knowledge engineering 860
  layer 43
  location-based services 857
  matchmaking 50
  service discovery 44
Semantic Web 14, 31, 375, 380, 856, 862
  Rule Language (SWRL) 858
semantics 377
semiotic
  level 11, 27
  tool 327
semiotics 14, 31, 380
sending window 788
sensing layer 865
sensor 125, 1035
  data 126
  fusion 126
  information fusion 863
  layer 864
  MAC (S-MAC) 328
  model 1034
  network 276
sensory system 651, 653
serializability 233
sequence diagram 716
serial port profile (SPP) 344, 1014
Series 60 249
server 219
  /client paradigm (SCP) 954
  -side handheld
    computing 302
    programming 309
  connection 251
  site 790
service 774, 884, 986
  -oriented
    approach (SOA) 529
    architecture (SOA) 71, 874, 878
    computing (SOC) 71, 626, 877–878, 884
    paradigm 621, 883
    pervasive computing 883
  abstraction layer pattern 746
  access control 800
  advertisement 884
  aggregator 7
  client 878, 884
  configurator 747
  creation environment 873
  delivery platform (SDP) 870
  discovery 53, 89
    application profile (SDAP) 344
    database (SDDB) 882
    protocol (SDP) 50, 272
  level agreement (SLA) 383
  location protocol (SLP) 43
  manager (SM) 799, 803
  plane 873
  provider (SP) 714, 878, 884, 982
  redundancy 53
  registry 878, 884
  relationship management (SRM) 873
services
  composer 86
  deployer 86
  discovery engine 86
session
  description protocol (SDP) 590
  initialization
    key (SIK) 1021
    protocol 590
  initiation protocol (SIP) 140–142, 595, 803, 873, 966
  key 1021
  mobility 967
SGML (see Standard Generalized Markup Language)
shared key authentication 1030
shooting Game 929
short
  message service (SMS) 32, 194, 252, 456, 526, 568 , 643, 666, 668, 726, 755, 757, 840, 885, 933
  sound file 670
shot boundary detection 889–893, 890
SI (see semantic interpretation)
sign-off/check-out mode 234
signal
  -to-noise ratio (SNR) 16, 595
  attenuation (SA) 511
signature
  attack 828
  scanner 651
SIK (see session initialization key)
SIM card 585
simple
  message service (SMS) 352, 356, 616, 620
  network management protocol (SNMP) 1026
  object access protocol (SOAP) 376, 795
simplest path algorithm 857
simulation
  -based learning 901
  game 928
simultaneous PoC session 727
single-user game 929
SIP (see session initiation protocol)
site
  -specific service 580

  design 285, 290
  transaction manager (STM) 949
SLA (see service level agreement)
SLP (see service location protocol)
SM (see service manager)
SMA (see stationary monitoring agent)
small
  mobile host (SMH)
  screen devices 123
smart
  antenna 942, 946
  card 472, 475, 480, 708
  cellular phone 308
  phone 327, 344, 757, 894, 898, 905, 1014
  space 71, 124
smartphone 304
SME (see station management entity)
  transaction manager (STM) 949
SMH (see small mobile host)
SMIL (see Synchronized Multimedia Integration Language)
SMS (see short message service)
SMS (see simple message service)
snap-on/wireless keyboard 571
sniffing 1022
SNR (see signal-to-noise ratio)
SOA (see service-oriented approach)
SOA (see service-oriented architecture)
SOAP (see simple object access protocol)
SOC (see service-oriented computing)
social
  context 116
  research 557
  system 632
software
  agent 340
  component 61
  development toolkit (SDK) 12, 394
solidring structure 822, 825
sophisticated service discovery protocols (SDPs) 43
sound 669–674
source
  -specific multicast (SSM) 541, 592, 595
  system 947
SP (see service provider)
space division multiple access (SDMA) 944
SPAM 888
spamming 534
spatial
  diversity 165
  models and ontologies 857
  services 792
speaker verification (SV) 677
spectrum management 730
speech 645
  application language tag (SALT) 647, 1053, 1056
  recognition 675
    grammar specification (SRGS) 1053–1054
Speech Synthesis Markup Language (SSML) 1056
SpeechPAK TALKS 570
SPIM 888
split mode 948
sponsoring 552
spontaneous service emergence paradigm 957
sports game 929
SPP (see serial port profile)
spyware 558, 561
SQL (see Structural Query Language)
SQL (see Structured Query Language)
SRGS (see speech recognition grammar specification)
SRM (see service relationship management)
SSL (see secure socket layer)
SSM (see source specific multicast)
standard
  transmission protocol 160
  type phone bill 552
Standard Generalized Markup Language (SGML) 142
stateful address autoconfiguration 258
stateless address autoconfiguration 258
static
  access frequency (SAF) 751
  agent 218
stationary monitoring agent (SMA) 328, 330, 333
station management entity (SME) 196
STM (see site transaction manager)
storage 536
  cost 231
  efficiency 231, 836
  inefficiency 227
  protection 840
stream control transmission protocol 813
streaming 77, 266
Structured Query Language (SQL) 370, 533
structured
  non-super peer 736
  super peer 736
student support 901
subscriber identity module (SIM) 840
  card 839
summarization technique 121
super
  frame 927
  peers 734
superior healthcare delivery 1010
supervised agent transport 430
sustainable
  agriculture 763
  development 763

SVC (see scalable video coding)
SVG (see scalable vector graphics)
SVG Basic (SVGB) 1050
SVGB (see SVG Basic)
SVG Tiny (SVGT) 1050
SVGT (see SVG Tiny)
switching cost 145
SWRL (see Semantic Web Rule Language)
Symbian 342, 503
  operating system (OS) 249, 304, 308, 895
symbolic model 857
synchronization 203, 206, 757
Synchronized Multimedia Integration Language (SMIL) 376, 1056
synchronous connection oriented (SCO) 344
syntactic translation 123
system
  architecture 1012
  capacity 946
  design 352, 616
  module 342

T

t-test 897
tablet PC 755, 757
tacitness 523
tagged image file format (TIFF) 670
talk burst control 727
Talks 570
TAM (see technology acceptance model)
tangible score 336
target respondents 640
task-orientated agent 578
Tatoes authoring 321
taxonomy 723
TCP-friendly rate control (TFRC) 20
TD-SCDMA (TD-SCDMA) 946
TDD (see time division duplex)
TDMA (see time division multiple access)
TDOA (see time difference of arrival)
teacher support 901
team (or guild) game 929
technical heterogeneity 399
technobabble 568
technology
  -conditioned approach to language change and use (TeLCU) 562, 568
  acceptance model (TAM) 38, 296–297, 894
  readiness 484
TeLCU (see technology-conditioned approach to language change and use)
tele-worker 634
telemedicine 504
telephone interview 887
teleradiology 540
temporal logic 995
temporary disconnection 75
terminal equipment 7
testing 12, 28
texting 757
text messaging 899
text to speech (TTS) 645–646, 1053
TFRC (see TCP-friendly rate control)
TFT (see thin film transistor)
TFTP (see trivial file transport protocol)
thematic spatial information services interfaces (TSISI) 790
thin film transistor (TFT) 536
third generation (3G) 475, 503, 533, 611, 933
  game 187
  mobile
    network (3G) 733
    system 682
  network 778
  wireless system 96
third generation (3G)
three
  -dimensional (3D) animation 672
  -tier Web-based architecture 1010
thumb board text interface 755
tight coupling 359, 810
time
  and distance sensitive (TDS) 751
  division duplex (TDD) 272, 940, 946
    -synchronous code division multi-access (TD-SCDMA) 946
  difference of arrival (TDOA) 511
  division multiple access (TDMA) 329, 456, 535
  of arrival (TOA) 511
timestamp 740
TOA (see time of arrival)
Toku number 636
Tokusuru
  Information Board 638
  Menu 638
tools 12, 28
tourism 365, 392
tourist
  guide 713
  services 660
TPI (see transport information items)
tracking 212, 660
  system 207, 907
traditional shortest path algorithm 857
traffic
  analysis 1024
  control 660, 730
transcoding 628–629, 984

transmission error 192
transparency 27
transport
  information item (TPI) 786
  layer mobility 968
  management 660
traveling mobile agent 579
tree
  -based routing 395
  structure 821, 825
    (see also hierarchical structure)
trialability 413–414, 418
triangle routing 259
triangulation 394
trilateration 770
triple play 803
trivial file transport protocol (TFTP) 263
trusted checkout 312
trusting agent 110
trustworthiness 25
TSISI (see thematic spatial information services interfaces)
TTS (see text-to-speech)
tunnel
  entry point 255
  exit point 255
tunneled IPv6 packet 255
tunneling 255, 543
tuple forming 865

U

u-commerce 310–316
ubiquitous
  computing (UC) 62, 561, 589, 775, 954
  tracking system 208
ubiquity 525
UC (see ubiquitous computing)
UCA (see uniform circle array)
UDDI (see universal description, discovery and integration)
UE (see user equipment)
UI (see user interface)
UID (see user interface design)
ultra wide band (UWB) 427
ULA (see uniform linear array)
ultrasonic signal 864
UMA (see unlicensed mobile access)
UMTS (see universal mobile telecommunication system)
unaccounted attribute 334
Unicode Standard 376
Unified Modeling Language (UML) 716
unified identity 140
uniform
  circle array (UCA) 942
  linear array (ULA) 942
  resource identifier (URI) 376
uniframe 212
  resource discovery system (URDS) 212
universal
  description, discovery and integration (UDDI) 89
  mobile telecommunication (UMTS) 535
  mobile telecommunications system (UMTS) 20, 194, 511
  plug and play (UPnP) 882
  serial bus (USB) 368
  SIM (USIM) 840
universally unique identifier (UUID) 44
unlicensed mobile access (UMA) 142, 360–361, 701
UNO (see user navigation ontology)
unsolicited message 399
unstructured super peer 735
unsupervised agent transport 432
update broadcast 751
updating
  distributed key 1019
  session keys 1018
uplink synchronization 944
UPnP (see universal plug and play)
URDS (see uniframe resource discovery system)
URI (see uniform resource identifier)
usability 27, 118
USB (see universal serial bus port)
user
  -centered operability 140
  -oriented rekeying 834
  attention model 599, 604
  context 26, 856
  control 800
  datagram protocol (UDP) 583
  equipment
  equipment (UE) 63, 727, 782, 942
  identification 775
  interface 303
  interface (UI) 74, 342–344
    design (UID) 319, 644
  mobility 967
  navigation ontology (UNO) 858
  preference 602
  profile 31, 126, 380, 626, 629–630
  terminal (UT) 810
utility function 44, 168
UUID (see universally unique identifier)
UWB (see ultra wide band)

V

VAG (see virtual anycast group)
validation authority (VA) 582
value 1004

  -added mobile service 497
  chain model 497
values and goals readiness 485
VANET (see vehicular ad hoc network)
VC (see virtual community)
VCEG (see Video Coding Experts Group)
VCoIP (see videoconferencing over IP)
vector space model (VSM) 87
vehicular ad hoc network (VANET) 424, 427
verification 430
Verisign 35
vertical
  handoff 151
  handover 967
    management 812
  mobile business 442
very large scale integration (VLSI) 427
video
  -on-demand 890
  coding 590
  sequence 889–893
  stream 124
  transcoding 627–631
Video Coding Experts Group (VCEG) 595
videoconference 758
videoconferencing 589
  over IP (VCoIP) 589
video telephony 589
viral marketing 98, 402
virtual
  anycast group (VAG) 54
  cluster 333
  community (VC) 632, 634
  component pattern 746
  machine (VM) 996, 998
    monitor (VMM) 996, 998
  memo 714
  private network (VPN) 533, 842, 1032
  proxy pattern 746
  reality (VR) 672
Virtual Reality Modeling Language (VRML) 672, 1052
virtualization 998
virus 1024
  attack 534
visibility 413–414, 418
visual query language 370
vitals 951
VLSI (see very large scale integration)
VM (see virtual machine)
VMM (see virtual machine monitor)
voice
  -activated communication 571
  -over Internet protocol (VoIP) 138, 142, 269, 700, 735, 898
  recognition 999–1003
Voice Extensible Markup Language (VoiceXML) 1050
VoiceXML (see Voice Extensible Markup Language) 1050, 1056
VoIP (see voice-over Internet protocol)
VPN (see virtual private network)
VRML (see Virtual Reality Modeling Language)
VSM (see vector space model)

W

W3C (see World Wide Web Consortium)
W3C-MMI (see W3C--MultiModal Interaction)
W3C-MultiModal Interaction (W3C-MMI)
WAN (see wide area network)
WAP (see wireless application protocol)
WAP identity module (WIM) 585
war-driving 1023
water-filling principle 169
WAVE file 669
WB (see wideband)
WCDMA (see Wideband Code Division Multiple Access)
Web
  feature service (WFS) 133, 137
  GIS 219
  map service (WMS) 133, 137
  Ontology Language (OWL) 377
  page 877
  service 77, 84, 89,795
    -based LBS (WS-LBS) 789
  Services Description Language (WSDL) 89, 795
  system 1050
  technology 536
Webtop client 80
WFS (see Web feature service)
Wi-Fi (see wireless fidelity)
  access point 701
wide area 207
  network (WAN) 253, 533
wideband (WB) 458
  code division multiple access (WCDMA) 534
widgets 81
WiMAX (see worldwide interoperability for microwave access)
WiMobile 1042
WinCE (see Windows Consumer Electronics or Windows CE)
Windows
  -based smartphone 305
  Consumer Electronics or Windows CE (WinCE) 793, 795
  Mobile 304, 309
wired
  equivalent privacy 1030
  Internet 298, 300

Wireless
  Application Protocol 2.0 (WAP 2.0) 423
  Markup Language (WML) 191, 194, 376, 456, 476, 583
wireless 283, 285, 327, 1000
  access network 391
  advertising
    (see also mobile advertising, mobile marketing)
  application protocol (WAP) 190, 194, 252, 376, 476, 480, 503, 913
    communication 744, 947
    connectivity 757
  communication network 195
  datagram protocol (WDP) 584
  data network 165–171
  device 62
  fidelity (Wi-Fi) 427, 507, 733
  headset 4
  healthcare 1010
  hotspots 207
  identity module (WIM) 480
  Internet 298, 472
    connectivity 576
  local area network (WLAN) 702, 733, 839, 856, 906
  local area networking (WLAN) 62, 139, 394, 455, 50 5, 507, 511
  mesh network (WMN) 921
  multimedia application 195
  network 62, 456, 832, 906, 1022, 1029
    architecture 1038
  personal area network (WPAN) 272, 276
  public key infrastructure (WPKI) 585
  routing algorithm 925
  scheduling mechanism 924
  security 1028–1033
  sensor 4
    network (WSN) 202, 206, 328, 424, 427, 921, 1034–1037
  session protocol (WSP) 190
  technologies 253
  transaction protocol (WTP) 583–584
  transmission 627
  transport layer security (WTLS) 584
  user behavior 476
  Web service 476, 480
WISDOM 573
WLAN (see wireless local area network)
WML (see Wireless Markup Language)
WMLSCrypt 585
WMN (see wireless mesh network)
WMS (see Web map service)
word of mouth (WOM) 300
workflow 1045
  management 1043–1049
workload 626
World Wide Web Consortium (W3C) 379, 463, 671, 675, 1050
worldwide interoperability for microwave access (WiMAX) 733
WPAN (see wireless personal area network)
WPP (see WAP peer protocol)

WS-LBS (see Web service-based LBS)
WSDL (see Web Services Description Language)
WSN (see wireless sensor network)

X

XDMC (see XML document management client)

XHTML (see eXtensible HyperText Markup Language)
  -MP (see XHTML Mobile Profile)
  Basic 376
  Mobile Profile (XHTML-MP) 423
XLink (see XML Linking Language)
XML (see eXtensible Markup Language)
  Linking Language (XLink) 377
XP (see extended profile)
XSLT (see eXtensible Stylesheet Language Transformations)
Xtensible Stylesheet Language Transformations (XSLT) 377

Z

ZigBee 273, 428, 507, 805
Zire 489
Zone Cooperative (ZC) cache 173–174